Eötvös Loránd University of Science
PhD School of Informatics

Balázs Szalkai

Algorithmic Problems In Bioinformatics

PhD Thesis

Eötvös Loránd University of Science
Doctoral School of Informatics (head: Prof. Dr. Erzsébet Csuhaj Varjú)
Program of Information Systems (head: Prof. Dr. András Benczúr)

Supervisor: Prof. Vince Grolmusz
Department of Computer Science
Eötvös Loránd University of Science

Budapest, 2018.

Acknowledgements

Hereby I would like to thank my supervisor Vince Grolmusz for helping me study a multitude of interesting topics in Bioinformatics which led us to a wide range of discoveries. I would also like to thank fellow PhD students Csaba Kerepesi and Bálint Varga for their cooperation in research requiring teamwork.

I would also like to thank all my co-authors for their valuable contributions:

Dániel Bánky
Vince Kornél Grolmusz
Kinga Nagy
Ildikó Scheer
Beáta Vértessy

Contents

1 Introduction

1.1 Motivation

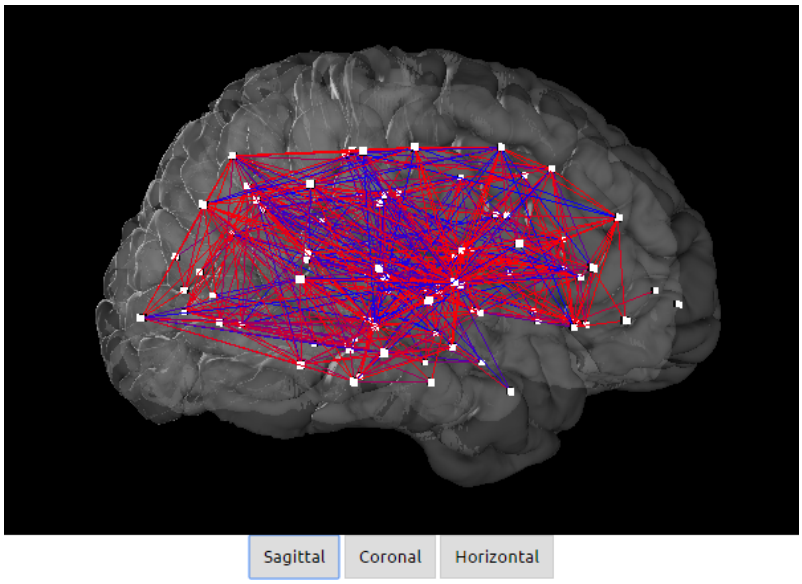

Figure 1: Screenshot of the Budapest Reference Connectome server

Biological research often results in big datasets which are then intended to be processed by the power of computers. These "big data" can be of various nature. We are speaking of data tables containing results of drug trials for thousands of people, collections of as much as several Gbp (giga-base-pairs) of DNA sequences, MRI images, protein–protein interaction graphs, 3D structural data of proteins, and neurological networks describing the micro- or macroscale structural connectivity of a part or whole of the nervous system of an organism.

A relatively well-known example is the "brain graph" of the nematode *Caenorhabditis elegans* [151] which has been subject to numerous studies and simulation attempts like the OpenWorm [143] project. Though this graph consists of only a few hundred neurons and synapses of this simple worm, it is still beyond human comprehension at the moment and requires computational methods for us to infer its properties. We may also compile graphs of the human brain, though here the vertices do not correspond to neurons but to larger areas called ROIs (regions of interest).

Without efficient computer algorithms, analysis of these data would certainly be impossible.

These days one can easily find a lot of bioinformatical data on the internet. These data are either open access or accessible on request. An unfathomable amount of resources have been invested in collecting these data, but we believe (and demonstrate) that not all information has been retrieved from these datasets by those who have compiled them. In other words, one can make additional exciting discoveries by using existing datasets made available by other researchers. The databases have usually been compiled for one specific purpose, to test a given hypothesis (e.g. the efficacy of a new medication), but we think that meaningful research can be done just by looking at these datasets from a different perspective, and that is what being a "bioinformatician" means to me.

1.2 Outline

The first section of this thesis describes our generalization of the k-means clustering algorithm. The original k-means algorithm only works when the data points are in a Euclidean space. This is a very important limitation, as more often the data points are elements of some abstract space, and a non-Euclidean distance metric is defined on them. We showed that the classical k-means algorithm can be generalized to non-Euclidean scenarios as well, when we have an arbitrary distance matrix, which does not even have to describe a metric on the data points. Our algorithm has since been applied by numerous other researchers [8, 22, 63]. Section 2 is based on our articles [133, 134].

One of my first results was using a drug trial database to infer association rules for Alzheimer's disease (AD). This disease is a great burden to advanced societies, where people tend to live comparably long thanks to today's advance medicine. While now we easily live into our 80's or even 90's, we cannot really escape the various kinds of dementia that go along with increased longevity, including Alzheimer's [24, 35]. We do have some drugs for AD that slow down the disease progress, but in most cases

the diagnosis is made too late, with a great proportion of brain neurons already dead. Thus we wanted to find biomarkers for Alzheimer's which could predict the disease well in advance.

We used a database by CAMD (Coalition Against Major Diseases) [121], which contained subject data for 11 drug trials, with demographical data, blood panel results, mental health questionnaires and cognitive test scores. Our goal was to find combinatorial association rules about Alzheimer's and dementia in general. That is, we searched for logical implications where the left side is an AND/OR combination of attribute-value equalities, and the right side is somehow connected to dementia. For example, the following expression counts as a combinatorial association rule:

$$sodium = high \wedge (protein = high \vee age \geq 60) \implies mmse\_total \leq 15 \qquad (1.2.1)$$

The program we have developed used exhaustive search to generate expressions like this and investigate the "truthfulness" of these rules. The results are interesting by themselves as well, but we also have deployed this newly developed data mining program on the internet, both in downloadable and installable form and as an open access webserver. This webserver is named SCARF, which is an acronym for Simple Combinatorial Association Rule Finder. This allows everyone to do combinatorial association rule mining without having to install and learn to use an offline utility. Section 3 describes these results and is based on our article [136].

According to my experience, the more easy to use a bioinformatics software is, the more popular it tends to become among researchers. This is not surprising, since a lot of users are biologists with limited programming experience, maybe having a different mindset than that of programmers, and thus a user interface designed by a programmer for programmers is not always suitable for them. An excellent example is the case of webservers. Online applications are often preferable to installable programs, because sometimes the most difficult part is to download a software, compile it, set it up and get it working. As easy as it sounds, it can be a tedious

work, and that is exactly why we developed e.g. the AmphoraNet webserver [91], which is precisely an online version of the AMPHORA2 pipeline, which is a program that previously had to be installed by the users on their computer and used offline. The AmphoraNet interface contains only the most important options so as it does not confuse the user, who is now free of the burden of system administrator tasks. I believe that, along with their truly superior performance, this user-friendliness contributed greatly to the popularity of the well-known bioinformatics tools BLAST (Basic Local Alignment Search Tool) [9] and MG-RAST [153], among others. That is why I felt it crucial to deploy SCARF as an online webserver, too.

Sequencing the whole human genome was undoubtedly a milestone in the history in bioinformatics. The Human Genome Project, launched specifically for this purpose, ran from 1990 to 2003 and cost 3 billion USD. Nowadays whole genome sequencing (WGS) costs less than $10,000 per sample, and is already close to the $1000 mark.

Next generation sequencing methods provide us with a vast amount of data. The question is, what can we do with all these data? The major challenge seems no longer reducing the cost of sequencing, nor improving the speed of sequence assembly and post-processing, but finding new applications to this already highly efficient process. Whole genome sequencing is used to discover links between mutations in human genomes and diseases, diagnose heritable conditions, and find previously unknown but useful genes in bacteria and archaea.

A similar technology brought the scientific field of metagenomics to life [40]. Before the advent of next generation sequencing, determining the bacterial composition of an environmental sample was done by culturing the organisms, then counting the colonies or the individual bacteria under a microscope. Though this process is simple and has a low material cost, it is slow because it requires human intervention and cannot be automated easily. But the main problem is that only a fraction of the organisms will be readily cultured in a laboratory, namely those bacteria and archaea that thrive on the selected substrate. For example, extremophiles will not be able to survive in classical laboratory conditions, exactly because they need extreme

environments to survive. In addition, this method is not suitable for the discovery of unknown organisms because we do not know how to culture them.

To escape from this vicious cycle, we can use next generation sequencing methods to obtain unbiased information from environmental samples. The sample is first handled so that its DNA content (also called the *metagenome*) is extracted, then the DNA is fragmented to yield pieces of a few hundred base pairs. Then these short sequences are processed further using biochemical methods like PCR, and then placed under a sensor array which detects the order of nucleotides in a given short sequence. The resulting data consists of millions of these short reads. The great advantage of this approach is that the short reads are taken randomly from the remains of the organisms originally living in the sample, and, while the results are only a small fraction of the original DNA material, enough of them are sampled, and each DNA fragment is equally likely to be included in the resulting dataset. In other words, it is a statistically correct metagenome sampling method.

The resulting dataset can then be used to infer the properties of the original environmental sample. The reads can be assigned to known species or taxonomic groups of microorganisms, and so the taxonomic composition of the environmental sample can be estimated from the metagenome [54, 165]. Care should be taken to account for the genome sizes of different bacteria, because an organism with a larger genome will have more reads sampled than an organism with a smaller genome.

Besides taxonomic analysis, reads can be aligned to a known database of protein coding genes, and the *functional composition* of the metagenome can also be determined. For example, we may discover that the organisms in a sample have an unusually large number of methane-related genes, and this may imply that they were adapted to the high hydrocarbon content the environment, which may mean that we can discover organisms in this environment suitable for decomposing oil in polluted areas.

We developed a new method called *the metagenomic telescope* [138] for determining the previously unknown function of genes in higher organisms. We first constructed a HMM (Hidden Markov Model) [23] on known DNA-repair genes. Then we

searched for DNA repair genes of microorganisms living in extreme environments, using this model. We chose metagenomes of an acidic mine drainage, hotwater springs and a wastewater plant [124] because we assumed that organisms living in these extreme conditions will have better DNA repair mechanisms. After that, we collected both the results and the original sequences, and constructed an enriched HMM from all these data. Then we used this enriched HMM alongside the original HMM on the genomes of higher organisms (human, dog, chicken, beef, etc.) to find genes which likely code proteins involved in DNA repair. The enriched HMMs found more sequences than the original ones. By examining the 3D structure of the proteins encoded by the newly found genes, we could indeed prove in some cases that the proteins encoded by the newly characterized genes are not only sequentially similar to other DNA repair proteins, but they have a similar structure and form similar multimers. This could mean that they are indeed involved in DNA repair, or at least they have evolved from those kinds of proteins. This demonstrated the power of the metagenomical telescope. To sum up, this method means enriching a database of related sequences with close matches from metagenomes, then using the new model to find more matches in higher organisms. Section 4 is based on our article [138].

One does not need to apply more complicated methods like HMMs for metagenome analysis. Simply counting the number of occurrences of a short nucleotide sequence can lead to interesting new discoveries. For example, it is already widely known that the GC-content (the number of G or C nucleotides) of the genome is specific for bacteria species [25, 79, 130]. But what if we counted sequences longer than a single nucleotide? Perhaps we could classify genomes or metagenomes based on the frequency of sequences that are a few nucleotides long. Of course, sequences with about 1000 nucleotides are definitely specific for genomes or metagenomes, since a sequence of this length may correspond to a gene encoding a protein consisting of about 333 amino acids. These are the two extreme cases: we know that one-nucleotide sequences like G and C are meaningful, and we also know that sequences of several hundred nucleotides are also meaningful. But even sequences of length

50 can be genome-specific: several nucleotide markers of this length were already identified [146]. Nevertheless, the frequency of sequences of length about 10 has not yet been employed in genome or metagenome classification. However, we recently demonstrated that the frequency of certain sequences of length 9 in gut metagenomes may be associated with diabetes. Section 5 is based on our article [135].

My recent results show that the connectomes (i.e. brain graph) of females and males tend to show different graph theoretical properties [140]. Given a diffusion MRI image for a subject's brain, one can construct a graph from it which represents the connectedness of the various brain areas. First the brain is segmented into gray matter and white matter, based mainly on fractional anisotropy but also with anatomical considerations. Fractional anisotropy (FA) is a value between 0 and 1 which describes the main diffusion directions at a voxel: if there is strong diffusion along an axis but almost no diffusion in other directions, then it is close to 1; and if there is equal diffusion in all directions, it is 0. Gray matter on average has a smaller FA than white matter, as white matter is made out of axons where water diffuses mainly in the direction of the axon. After the brain has been segmented into gray and white matter, the parcellation algorithm is run to identify the ROIs (regions of interest), by trying to morph the image onto a reference brain which has previously been parcellated by hand. The tractography algorithm is run in parallel: this module is responsible for the tracing of axons. By combining the results of parcellation and tractography we can construct a graph whose vertices are the ROIs and whose edges are the nerve tracts connecting these regions [42]. We can use multiple resolutions (brain atlases) at the parcellation stage, which means that the resulting brain graph can have less or more vertices. Choosing the right atlas for a connectome is always a compromise: atlases with fewer ROIs are anatomically more sensible, while atlases with more ROIs result in a larger graph which may have more interesting graph-theoretical properties. We used the Connectome Mapper Toolkit [42], which implements the pipeline described above, to get connectomes from MRI images. Then we examined the connectomes of women and men, and compared them to each other.

Although the connectomes of women and men turned out to be quite similar, we saw that the female connectome had a significantly larger connectivity: it had more edges in general, and also the proportion of inter-hemisphere edges was greater than that in males. By mathematical analysis we also concluded that the female connectome is a better expander and has more spanning trees. We have to emphasize that these differences are only about the *connection structure* of the brain and do not allow one to jump to conclusions about the *working* of the brain. Section 6.4 describes these findings in detail, and is based on our article [140].

We have also made available a 3D brain graph model online called the Budapest Reference Connectome [137], where one can view a consensus connectome, which means that we averaged the connectomes of several people into one graph, and that graph can be visualized in an interactive web application. See Section 6.3 for more information. This section is based on our article [137].

The subject database of the Human Connectome Project [103] also let us make interesting discoveries about the correlation of brain ROI sizes, psychological test scores and cognitive scores. This data table contained 527 rows (one for each subject in the study), and 451 columns (attributes). For each subject, a wide variety of data were available, including demographic data, results of cognitive and mental health tests like MMSE (Mini Mental State Exam), NIH toolbox [150] (psychological and cognitive tests), and NEO-FFI (a five-factor inventory for personality testing). A more interesting part of this database was the inclusion of the volume of brain regions obtained by parcellation: those at Human Connectome Project who compiled this database ran the parcellation utility of the FreeSurfer [56] software package to identify the ROIs of each brain, then calculated the volume of the subcortical regions, and the average thickness and area of the cortical regions (whose product equals to the volume anyway). We augmented this database by including the graph-theoretical parameters of the connectomes.

Our goal was to discover interesting correlations between the attributes. In other words, we aimed for selecting a set of attribute pairs with the most important correlation, which is exactly an undirected graph on the attribute set, whose edges

correspond to correlated attribute pairs. To achieve this, we calculated a maximum-weight spanning tree on the attributes, where the weight of each edge was the absolute correlation between the two attributes. The reason for this was that we wanted to uncover the hierarchical structure of the attributes, filter the information contained in the correlation matrix, and avoid cycles in the resulting graph because if X correlates with Y and Y correlates with Z, then it is "likely" that X will correlate with Z, too, and then the XZ correlation does not represent much new information. This maximum-weight spanning tree approach is similar to the one used by Mantegna et al. [101], who explored connections between stock daily returns.

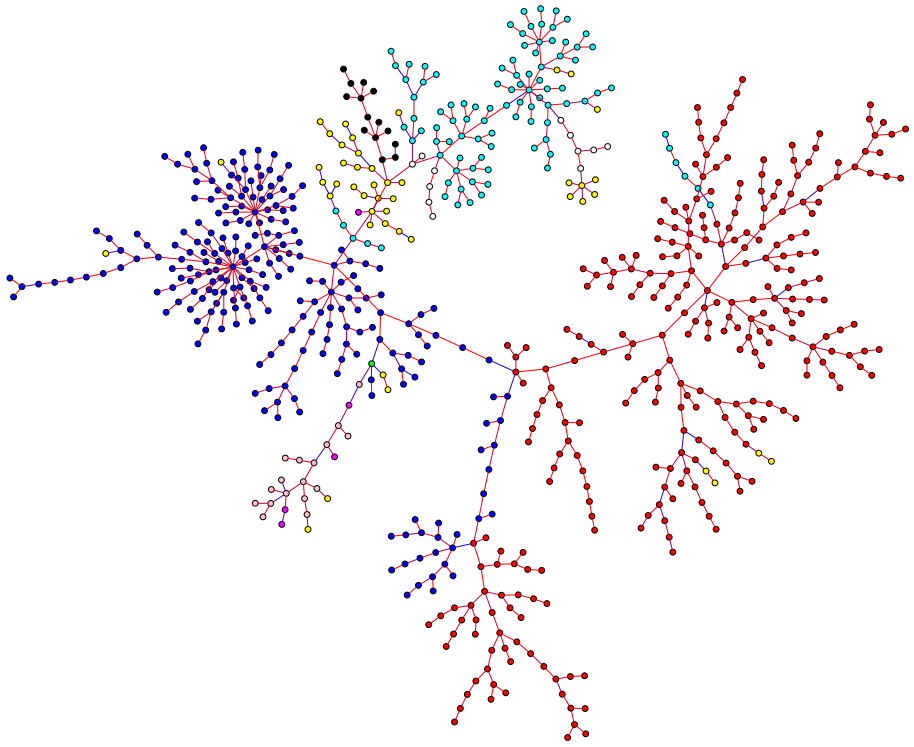

Figure 2: The correlation tree

The spanning tree showed some results we already expected, and some that were new and interesting discoveries to us. On one hand, the emotion scores of the NIH toolbox formed a subtree just as expected, but the way they became connected in the tree now allows us to cluster these attributes, and obtain a big picture on how

these emotions are correlated with each other. For example, stress was a major hub of this subtree, connected to life satisfaction, self-efficacy, sadness and anger-hostility; and the tree also shows that friendship is more connected to emotional support that instrumental support. The subtree corresponding to ROI sizes has two major hubs. It is probably not surprising that these hubs are the left and right hemisphere volume, which are dependent on the total brain volume and, in turn, determine the sizes of the individual ROIs. Section 7 describes these findings in detail and is based on our paper [142].

I hope that my research has contributed to exploiting the vast amount of big data for the needs of humanity. We need a better understanding of diseases and the human body to make people's lives better. I believe that using the enourmous amount of freely available data and looking at them from a different perspective can help us a lot in achieving this goal.

2 A generalization of the k-means algorithm for arbitrary distance matrices

2.1 Introduction

The k-means clustering algorithm works for data points in a Euclidean space only. Therefore, it is applicable only for datasets where there is a correspondence from the data points to points of a Euclidean space $\mathbb{R}^n$. A possible generalization is described in this section which would allow arbitrary data points with an arbitrary distance function. This is achieved by first showing how k-means can be used for scenarios when only the distances between the data points are known instead of their exact locations in $\mathbb{R}^n$. We named this algorithm *relational k-means* [133]. Then relational k-means is generalized for non-Euclidean scenarios, and a simple $C\sharp$ implementation [134] is mentioned.

The traditional k-means method [99] works with some data points $p_1, ...p_n \in \mathbb{R}^d$. A desired cluster count N must be given in advance as well. The method then attempts to arrange the points "well" into at most N clusters. The result is a function $\ell : \{p_i\}_{i=1}^n \to \{1, ...N\}$ which labels the points with the index of the corresponding cluster. We define the *value* of a clustering with the following measure:

$$\sum_{i=1}^{n} ||p_i - z_i||^2,$$

Here $S_i = \{p_j : \ell(p_j) = \ell(p_i)\}$ is the cluster p_i belongs to, and $z_i = \frac{1}{|S_i|} \sum_{p_j \in S_i} p_j$ is the i*th centroid* (that is, the average of the points in the cluster S_i). The lower the value, the better the clustering. In other words, we would like to minimize the sum of the squared centroid distances.

In its traditional form, k-means cannot be applied to arbitrary data with an arbitrary metric (like some sequences with a dissimilarity measure), since the value of a clustering cannot be determined without taking the average of the data points.

Abstract data points cannot be averaged like points represented in a Euclidean space. A possible solution would we to try mapping the data points to some vectors so that the distances between the resulting vectors are close to the original distances, then run k-means on the resulting approximate representation. This technique is called *multidimensional scaling* [29]. The main drawback of this approach is that the approximate Euclidean representation may have a substantially different distance matrix that the original one.

The standard k-means method described above has been successfully generalized in various ways [37] [46], but none of the generalizations pointed in the direction of non-Euclidean scenarios. On the other hand, a similar clustering method called *fuzzy c-means* has been generalized for non-Euclidean cases [69]. The generalized method is called Non-Euclidean Relational Fuzzy C-means *(NERF c-means)*. Our solution which generalizes k-means to non-Euclidean data points can be seen as a vast simplification of NERF c-means. Our method is described in the next sections along with a simple implementation.

2.2 Relational k-means

The first step of the generalization is to forget about the vectors representing the data points, and work only with the pairwise distances. That is, we would like k-means to work for those cases as well when the Euclidean data points are hidden from us, but we know their distance matrix. If we get a solution for this problem, it will be able to help us generalize k-means for arbitrary distance matrices.

So let us suppose that $A \in \mathbb{R}^{n \times n}$ is the *squared distance matrix* of some Euclidean data points $\{p_i\}_{i=1}^n$. That is, $A_{ij} = ||p_i - p_j||^2$. In order for k-means to work, the only thing we need is calculating the squared centroid distances. In other words, we have to calculate $||p_i - z_i||^2$ without knowing the vectors, relying only on our knowledge about the distances.

The $p_i - z_i$ distance vectors are special linear combinations of the p_j data points. The sum of coefficients in such a linear combination is always zero. We can write

$p_i - z_i$ in a more general way as $\sum_{j=1}^{n} \lambda_j p_j$ for some $\vec{\lambda} \in \mathbb{R}^n$ coefficient vector which is perpendicular to the constant one vector $\vec{1}$.

The key idea is that we can calculate the squared norm of such linear combinations, knowing only the pairwise square distance matrix A:

$$||\sum_{i=1}^{n} \lambda_i p_i||^2 = \sum_{i=1}^{n}\sum_{j=1}^{n} \lambda_i \lambda_j \langle p_i, p_j \rangle = -\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} \lambda_i \lambda_j ||p_i - p_j||^2 = -\frac{1}{2}\lambda^\top A \lambda,$$

Note that the above reformulation works only for these special linear combinations, i.e. $\sum_{i=1}^{n} \lambda_i$ must be zero.

The above transformation yields us a way to calculate the centroid distances. It tells us that we only need to compute a quadratic form. Even though calculating a quadratic form is a computationally intensive operation, now we have a method which allows running any k-means heuristic, knowing only the squared distance matrix A. Of course, the above centroid distance calculation method can be applied to other clustering schemes as well, like fuzzy ones.

2.3 Non-Euclidean matrices

Let e_i now denote the ith standard basis vector (that is, the vector whose coordinates are all zero except for the ith one, which is 1). For an index set $S \subset \{1, ...n\}$ let $\chi(S)$ denote its *characteristic vector*, i.e. $\chi(S) = \sum_{i \in S} e_i$. Now, for a cluster S, let $z_S := \frac{1}{|S|}\sum_{j \in S} p_j$ denote the centroid corresponding to S.

Note that, if $\lambda = \frac{1}{|S|}\chi(S) - e_i$, then the squared centroid distance $d^2(p_i, z_S) := -\frac{1}{2}\lambda^\top A \lambda$ is also defined for those cases when A is not a Euclidean squared distance matrix. So relational k-means can be applied for non-Euclidean cases as well, since it only uses the squared distance matrix, which can then be substituted for by an arbitrary matrix in the formulas. This way we defined the weighted average of abstract data points in a way. Namely, we have given a possible definition of distance between two such objects.

With the above generalization, any k-means algorithm can be adapted to abstract data points with a distance function. However, this remains a theoretical result if some distance matrix A does not behave well. That is, if $-\frac{1}{2}\lambda^\top A\lambda$ is negative for some weight vector λ perpendicular to $\vec{1}$, then the squared centroid distance corresponding to λ will be negative. This shows that completely arbitrary distance functions behave in a different way than Euclidean ones.

A possible solution for eliminating negative distances is altering A so that A_1 becomes negative definite, where we get A_1 by restricting the quadratic form A to the vectors perpendicular to $\vec{1}$. This must be done in a way that the original matrix is modified as little as possible. This kind of matrix correction is not compulsory as the algorithm works for any matrix, but it may be advisable when clustering real-world data.

In [69] a method called *β-spread transformation* was proposed. This involves increasing all the pairwise distances by the same amount (that is, adding $\beta(J - I)$ to the matrix, where J is the matrix full of ones, and I is the identity matrix) until we have a matrix which does not yield negative squared centroid distances. It is trivial that such a β exists. This approach has been utilized for fuzzy c-means and was reported to work well in real-world scenarios.

2.4 The algorithm

The most well-known k-means heuristic works as follows. We start from an initial clustering (e.g. a random one), then calculate the squared distance of each data point from each centroid, then assign each data point to the cluster whose centroid was the closest to that point. This classical algorithm can be easily rewritten for a distance matrix instead of Euclidean data points, since now we can define the distances from the abstract centroids even if all we have is the squared distance matrix.

It can be proven easily that the value of the current clustering decreases by each iteration in the Euclidean case. However, this is not the case in general. Our

experiments show that performing a reassignment iteration can in fact *increase* the value, thus making the clustering worse than the previous one. So the generalized algorithm should stop in those cases when the value of the clustering did not decrease by the last iteration, and then the last performed iteration should be undone.

Like the classical k-means heuristic, this algorithm is finite as well. The number of possible clusterings is finite, and we decrease the value in each step (except for the last one). The question is, how fast we can perform a single iteration. The naïve implementation of the iteration step takes $\mathcal{O}(Nn^3)$ time, since we have to calculate Nn quadratic forms (each data point must be matched with each cluster).

Observe that the quadratic forms we need to calculate in fact involve smaller matrices—matching a data point with the cluster S_i can be done by calculating a $|S_i| + 1$-dimensional quadratic form. This means that the runtime of a step can be easily reduced to $\mathcal{O}(n^3)$. However, this is still too much. We will show that actually $\mathcal{O}(n^2)$ time is enough for an iteration. Let us assume that the diagonal of A is zero—this is true for all sensible distance matrices.

Let q_{ij} denote the squared distance between p_i and the abstract centroid of S_j. Let us rewrite q_{ij} in another form:

$$q_{ij} = -\frac{1}{2}\left(\frac{1}{|S_j|}\sum_{k\in S_j} e_k - e_i\right)^\top A \left(\frac{1}{|S_j|}\sum_{k\in S_j} e_k - e_i\right) = -\frac{1}{2|S_j|^2}\sum_{a,b\in S_j} A_{ab} + \frac{1}{|S_j|}\sum_{k\in S_j} A_{ik}.$$

The above reformulation is the sum of two quantities. One of them depends only on j, while the other summand is a function of both i and j. We can calculate the first part in $|S_j|^2$ time for the cluster of p_j, then cache it. Thus calculating the first part for all clusters can be done in a total number of $\mathcal{O}(n^2)$ steps. The second part can be calculated in $|S_j|$ time for a pair (i, j), so if we fix i, it can be calculated in n time for all clusters, which means that the total calculation time for all data points and clusters is $\mathcal{O}(n^2)$. Together we have that $\mathcal{O}(n^2)$ time is enough to perform an iteration step. This is not much worse than the $\mathcal{O}(nN)$ time required by the classical

algorithm, but we have to acknowledge that our generalization still has a cost.

2.5 Implementation

To produce a reasonably good clustering, we need to perform multiple *attempts*. An attempt is a full run of the algorithm from some starting configuration. Since we get different results when starting from different initial clusterings, we have a chance to improve our sofar best solution by running the whole algorithm several times.

We propose the following optimization scheme: fix an integer $K > 0$, then run the algorithm repeatedly, starting from different configurations each time, until there is a streak of K runs without finding a better clustering. It can be easily shown that, for some real number $0 < p_K < 1$ there is at least p_K chance that, if we would choose a random clustering, it would be worse than the one the algorithm yielded. In particular, p_K rapidly tends to 1 if K tends to infinity.

If we have P processors, we can perform P individual attempts in parallel, since they do not depend on the results of each other. This means that, in practice, the algorithm can be fully parallelized. Since modern computers tend to have an increasing number of processors with limited performance, the algorithm scales well with the current trends of technological advance.

We implemented the above parallelized algorithm in $C\sharp$. We tested the implementation on a set of more than 1000 proteins with some dissimilarity measure. Setting K to 20 and N (the number of clusters) to 10, the program finished in under a minute in most of the cases. This shows that even this quite simple implementation of the relational k-means algorithm is fast enough for real world use. A C++ implementation could be about twice as fast according to our estimates.

3 Association Rule Mining and Alzheimer's Disease

3.1 Introduction

The concept of combinatorial biomarkers was conceived around 2010: it was noticed that simple biomarkers are often inadequate for recognizing and characterizing complex diseases. In this section we present an algorithmic search method for complex biomarkers which may predict or indicate Alzheimer's disease (AD) and other kinds of dementia.

We applied data mining techniques that are capable to uncover implication-like logical schemes with detailed quality scoring. Our program SCARF is capable of finding multi-factor relevant association rules automatically. The new SCARF program was applied for the Tucson, Arizona based Critical Path Institute's CAMD database, containing laboratory and cognitive test data for more than 6000 patients from the placebo arm of clinical trials of large pharmaceutical companies, and consequently, the data is much more reliable than numerous other databases for dementia. The results suggest connections between liver enzyme-, B12 vitamin -, sodium- and cholesterol levels and dementia, and also some hematologic parameter-levels and dementia.

Dementia is presently a major problem of high-income countries and also an increasing concern of low-income nations worldwide. Though sporadic before the age of 60, its occurrence is doubled by every five years of age thereafter [24, 35]. About 40 percent of the population over 90 are affected, and up to 20 percent of those between 75 and 84 suffer from this condition [113, 158]. The most common cause of dementia is Alzheimer's disease (AD). The earliest symptoms of AD include memory problems; disorientation in time and space; and difficulty with calculation, language, concentration and judgment. As the disease evolves, patients may develop severe behavioral abnormalities and may even become psychotic. In the final stages

of the disease the sufferers are incapable of self-care and become bed-bound, for years or even decades, up until their death.

The diagnosis of AD in the great majority of the cases is done by clinical criteria, using standardized questionnaires [105]. Generally accepted evidences show that neuropathological damage begins more than 20 years before those clinical signs [86], and by the time it is diagnosed, a large part of the neurons are already irreversibly lost.

In the latest years, a quite reliable diagnostic method emerged by the combination of cerebrospinal fluid analysis, clinical signs and neuroimaging techniques [48]. The method, however, is prohibitively expensive, is not an early warning-type biomarker, and thus does not seem to be applicable for wide-scale screening of the senior population.

Very recently, using the combination of usual clinical laboratory data, cognitive impairment questionnaires and blood-based proteomics assays was reported to reliably diagnose AD, without neuroimaging or cerebrospinal fluid assays [108, 109]. However, early warning biomarkers are still need to be found.

Our final objective is finding new combinatorial biomarkers for Alzheimer's disease. In this section we describe our results that may be used to reach this final goal; but presently we are able to show only that certain sets of laboratory data may make dementia (and not AD) more probable, and certain other sets may make dementia less probable.

There are several large databases on Alzheimer's disease available for researchers. The quality of their data obviously depends on the methodology of the research that produced the database in question. We believe that the most well-organized, strictly overseen and rigorously documented experiments are those conducted on behalf of large pharmaceutical companies in hospitals and clinics in phase 1, 2 and 3 drug trials. We base this assumption on the fact that the laws governing the introduction of new pharmaceuticals require a costly, lengthy, thorough and well-documented testing process. Unfortunately, the detailed results of those trials are seldom published

(especially those corresponding to unsuccessful drug trials), presumably since they may constitute valuable industrial secrets of the companies that ordered the trials.

In their Alzheimer's disease database the Tucson, Arizona based Critical Path Institute made available the results of the placebo arm of numerous multi-million dollar clinical trials conducted on behalf of large pharmacological companies [119–121]. The data of the placebo line of the trials does not contain proprietary information concerning the effects of the novel drugs under trial, but it does contain reliable, well-organized laboratory and cognitive test-data, presumably in much higher quality than other, larger, but perhaps less strictly conducted and controlled studies for AD.

Data used in our study have been obtained from the Coalition Against Major Diseases (CAMD) database [121]. In 2008, Critical Path Institute, in collaboration with the Engelberg Center for Health Care Reform at the Brookings Institution, formed the Coalition Against Major Diseases (CAMD). The Coalition brings together patient groups, biopharmaceutical companies, and scientists from academia, the US Food and Drug Administration (FDA), the European Medicines Agency (EMA), the National Institute of Neurological Disorders and Stroke (NINDS), and the National Institute on Aging (NIA). Coalition Against Major Diseases (CAMD) includes over 200 scientists from member and non-member organizations. The data available in the CAMD database have been volunteered by both CAMD member companies and non-member organizations.

In contrast with more traditional statistical methods, we applied data mining techniques for data analysis and combinatorial biomarker search. Data mining, as defined in [68], is the analysis of large observational sets of data for finding new, still unsuspected relations with novel, usually high-throughput methods. Frequently, data mining uses large data sets originally collected for uses other than the data mining analysis [68].

Simple biomarkers (e.g., the high level of glucose in diabetes) show a physiological condition, related to the appearance or the status of a disease. The concept of combinatorial biomarkers appeared around 2010. Numerous authors use the term in

the following sense: If—say—the high concentration of all the molecules A, B and C characterizes a certain condition X well (and the high concentration of any subset of the set $\{A, B, C\}$ would not), then they say that $\{A, B, C\}$ is a combinatorial biomarker of the condition X [163]. In [109], by applying proteomics assays, a 30-protein set was identified as a combinatorial biomarker of AD.

We intend to discover more involved combinatorial biomarkers that may contain clinical laboratory data and psychiatric test data, and we count not only on positive findings (i.e., high concentration or appearance of a certain value), but also the lack thereof (i.e., normal or low concentration). We start with frequent itemset analysis, then apply association rule mining [68]. We apply a new methodology that discovers complex combinatorial biomarkers only if these complex biomarkers have stronger implications than the simpler biomarkers.

Therefore, our program SCARF will not produce artificially complex biomarkers: the more complex the new biomarker, the more valid the new implication.

3.2 Association Rule Mining

Our research group was among the first applying association rule mining in molecular biology [85]. Recently, association rule mining has been gaining applications in drug discovery [58], in the design of clinical trials [51], and most recently, also in image analysis in Alzheimer's research [36].

Association rule mining is a field of data mining [68] developed by marketing experts for discovering implication-like rules in uncovering customer behavior [6], without *a priori* assumptions on this behaviour. We intended to apply this method for laboratory and cognitive test data from the CAMD database [121].

We analyzed how the presence/absence/severity of cognitive impairment could be detected from combinations of known biomarkers, demographic information and measurements of vital signs. As an example, consider this expression:

$$sodium = high \ \& \ (protein = high \text{ or } age \geq 60) \rightarrow mmse\_total \leq 15 \qquad (3.2.1)$$

Here & stands for logical AND. This rule states that if blood sodium is high, AND urine protein is high OR age is at least 60, then the total MMSE (Mini Mental State Examination) score will be at most 15 out of 30. Let us call the left-hand side of the expression (abbreviated by LHS) a combinatorial marker of the right-hand side (abbreviated by RHS). Thus the statement above can be reformulated as follows: high serum sodium combined with either high urine protein or age of at least 60 is a marker of a total MMSE score less than or equal to 15.

An expression consists of *elementary clauses* combined by logical operators. These elementary clauses may include equalities and inequalities. By substituting all elementary clauses with some wildcard, we can obtain the *pattern* of an expression. For example, the expression above has the following pattern:

$$\square \ \& \ (\square \text{ or } \square) \rightarrow \square \qquad (3.2.2)$$

During our analysis we started with a given pattern like the one above. Then we considered all the possible logical expressions fitting this pattern, and assigned numerical values to them that indicated the reliability and validity of the logical rules. Then we filtered and sorted the vast amount of possible rules according to these numerical criteria, and selected the best ones. We changed a simpler rule to a more complex rule only if the more complex rule had higher reliability/validity than the simpler rule (see the next section for the exact definitions).

3.3 The CAMD database

Our data source, which will be referred to as CAMD from now on [121], was provided by the Coalition Against Major Diseases, and consisted of the placebo arm of several drug trials. Over 6000 subjects participated in these trials including demented

and not demented people of various age and sex (see Table 6) for basic statistics). Standard laboratory data that have been collected for the subjects included about 300 different values in blood or urine altogether. These values were generally measured multiple times per subject (on different visit days), though each person was tested for only about 30 different values. The cognitive and psychological status of the subjects was measured at different times by standardized questionnaires ADAS-COG, ADCS-ADL, MMSE, NPI and SIB. In addition, some genetic tests have been performed, e.g., ApoE and MTHFR genotypes were recorded. Vital sign measurements (BP, pulse rate, respiratory rate and body temperature) have also been taken. Results concerning this dataset will be described in greater detail below.

We transformed this large dataset into a conveniently processable form. The CAMD database contained several rows describing one person and these were scattered between multiple data tables. So we collected the essential data from CAMD into one single table: this simplified table contained only one row for each subject.

If a subject was tested on different visit days, then we took the average of these test results. The resulting main table for CAMD consisted of around 170 columns (record fields) and 6000 rows (entries).

3.4 Description of the algorithm

Our main method of processing the resulting table was association rule mining. First, we took a given pattern like $\Box$ & ($\Box$ or $\Box$) $\rightarrow \Box$. Notice that the LHS (Left Hand Side) is in conjunctive normal form here (multiple OR clauses ANDed together). This pattern can be encoded as "1 2", as the first OR clause has one sub-clause and the second one has two. This pattern matches all statements of the following kind: "if property A is present and property B or property C is present, then property D is present".

Since we are interested in implication-like association rules that indicate factors implying normal or demented mental state, we made restrictions on which data columns can occur on the LHS (Left Hand Side) and the RHS (Right Hand Side).

Laboratory data and sex were allowed on the LHS, and columns directly indicating mental status on the RHS. Then we gave numerical constraints on the "goodness" of a rule—thus introducing an ordering on the rules. Finally we tried to fill in all the void boxes in all possible ways to find the best rules.

If done without any optimization, this process would have yielded a vast amount of different rules that would have needed to be evaluated "by hand". Even just enumerating all the possible matches to this pattern would have required enormous computational resources. Consequently, we needed to make the computation feasible: we used a *branch-and-bound* approach similar to the Apriori Algorithm [68]: if certain values for the first two boxes made a rule fail our constraints—regardless of what would be written in the third box—, then we threw out the rule and did not bother checking all the possible values for the third box. (A good analogue would be cutting a tree in a clever way: one does not bother removing all the little twigs one by one, but rather cuts the trunk.) This technique saved us considerable computational time.

The association rule mining was done with our own program written in the C++ programming language, named SCARF (Simple Combinatorial Association Rule Finder). We calculated various standard numerical values for all association rules, which would indicate their validity. First, we defined the *universe* of a rule: this is the set of the database rows where all columns present in the rule have a known value. As we mentioned before, not all subjects were tested for everything, so our database contained a large amount of N/A entries. For testing the validity of a rule, only those rows could be taken into account, where there is no N/A written to any of the columns participating in the rule.

For evaluating the validity of a rule, we continued to work with only its universe and temporarily discarded all other rows in the database. Next, we calculated the *LHS support*, *RHS support* and *support* of a rule. The *LHS support* is the number of the rows where the LHS is true, the *RHS support* is the number of the rows where the RHS is true, and the *support* is the number of the rows where both the LHS and the RHS are true.

Then, we calculated the *confidence*, *lift*, *leverage* and χ^2-*statistic* for a rule. The *confidence* is defined as the conditional probability of the RHS, assuming that the LHS is true. If one has high serum sodium combined with high urine protein or age at least 60 in our example, then confidence describes the chance of having a low MMSE score. The *lift* shows how many times the presence of the LHS increases the probability of RHS. Generally it indicates how big a risk factor the LHS is—though it is not certain that the LHS *causes* the RHS, as they both may be only consequences of some background phenomenon [68].

The *leverage* is the difference between the observed probability of both the LHS and RHS being true, and the estimated probability we get by assuming that the LHS and RHS are independent events. It indicates the level of dependency between the LHS and the RHS in a way. Finally, the χ^2-*statistic* is a well-known measure of the estimated dependence of the indicator variables of the LHS and RHS. The *p-value* output by SCARF comes from this χ^2 test.

The *E-value* (also calculated by SCARF) equals to the p-value multiplied by the total number of possible rules, i.e. corrected for multiple comparisons. If we examine many rules, there is a high probability that the p-value will be small enough, while the E-value is insensitive for this kind of artifact.

The following table formalizes some of the above definitions. Here $\mathcal{P}$ denotes the probability measure:

$$\text{Confidence} = \mathcal{P}(RHS|LHS)$$

$$\text{Lift} = \frac{\mathcal{P}(RHS|LHS)}{\mathcal{P}(RHS)}$$

$$\text{Leverage} = \mathcal{P}(RHS \wedge LHS) - \mathcal{P}(RHS)\mathcal{P}(LHS)$$

For the CAMD database the acceptable values were set as follows: universe $\geq$

500, support ≥ 50, confidence ≥ 0.5, lift ≥ 1.2, p $-$ value ≤ 0.05. In particular, we recorded rules on data that were measured on at least 500 subjects. We defined the *goodness* of a rule to be equal to its lift.

Therefore we listed association rules of lift at least 1.2, i.e., only those rules were listed where the LHS increased the probability of RHS with at least 20%.

One of the most significant novelties in our approach was pruning (simplifying) those rules which were too complicated. The SCARF program threw out elementary clauses from the LHS as long as the overall goodness (i.e. the lift) of the rule did not decrease by more than 2%. Then it deleted the whole rule if its numerical values dropped below our constraints during the simplification process. In other words, we sacrificed some of the lift for simplicity, to avoid overfitting.

Having listed the best rules, we also tried to determine whether the elementary clauses (like $lb\_ast = h$, $lb\_folate = l$, etc.) have positive or negative effect on mental state. Therefore we counted their appearances on LHS, and classified these occurrences by the nature of the RHS: does it indicate normal cognition or rather dementia? We counted how many times an elementary clause occurred on the LHS of a rule when the RHS indicated a positive mental state, and how many times it occurred in rules where the RHS showed a negative state. Thus, in addition to mining rules whose LHS could probably serve as good combinatorial risk factor of dementia, we estimated the contribution of the *individual* clauses, for example "protein=$high$" to the onset of cognitive impairment.

For an elementary clause, *Positive score* was the number of rules with positive RHS, and *Negative score* was the number of rules with negative RHS. Then we compared *Positive score* with *Negative score* : by subtracting the negative score from the positive score we got a value called simply the *score* of the clause. Those elementary clauses whose score was positive were called *positive* clauses, and similarly, those where the score was negative were called *negative* clauses.

To summarize our method: we searched for combinatorial biomarkers using a branch-and-bound algorithm for association rule mining; then made statistical anal-

ysis regarding elementary clauses.

3.5 Results

The program output 725 rules from the CAMD database. Selected rules, ordered by lift (i.e. "goodness") decreasing are listed in Table 8. The whole set of rules is presented as Table S1 of the online supporting material.

On the LHS, clauses concerning biomarkers end in "=l", "=h", "=n", or combinations of these. Here l means low, h means high and n means normal. If there are multiple letters (such as nh), then the corresponding equality states that the value is either high or normal. In other words, single letters correspond to a value category, while multiple letters mean the union of these categories.

For example, the second rule in Table 8 was that of the second best lift. It can be interpreted in the following way: It is likely that if serum sodium level is elevated, and serum glucose level is either too low or normal, then the total MMSE score will be less than 15. Note that it is true for all rules of ours that there is not necessarily a causal relation between the LHS and RHS, as both the LHS and RHS can be consequences of an unknown process in the background.

The third rule states that "if serum sodium level is elevated, and calcium level is either low or normal, then MMSE orientation subscore will be at most 2". The seventh rule in Table 8 states that "if serum sodium level is elevated, and body temperature is too low, then total MMSE score will be less than 15".

From these selected rules we can conclude that elevated sodium combined with various other factors (not too high glucose, not too high calcium, low temperature) might be a good indicator (or even the cause) of mental decline.

Elementary clauses with the greatest positive effect on normal cognition are listed in Table 12.

Elementary clauses with the greatest negative effect on normal cognition are listed Table 14.

3.6 Discussion

Among the 725 rules identified, 513 had lift values exceeding 2.00. Most of the rules exceeding even the 3.00 lift value had one thing in common: the LHS contained the premise $lb\_sodium = h$.

3.6.1 Serum sodium

A great number of rules (224) have high sodium on the left hand side, all of which have impaired cognition on the right hand side. Net water loss is responsible for the majority of cases of hypernatremia [4]. A recent publication, examining the causes and comorbidities in patients older than 65 years, has found that the most common cause of community-acquired hypernatremia is dehydration due to reduced oral intake [147]. More interestingly, they found that the most common comorbidity in this patient group was AD, present in 31.4% of patients with hypernatremia [147]. Hydration status has a significant impact on the volume of grey and white matter in the brain and on the quantity of cerebrospinal fluid as a hallmark of ventricular enlargement [129]. The pattern of shrinkage in white matter volume and increase of the ventricular system due to dehydration is consistent with the structural brain changes observed during the progression of AD [129]. In another study, patients with AD underwent bioelectrical impedance vector analysis to assess the body cell mass and hydration status related to AD [33]. Results demonstrated a tendency towards dehydration in patients with AD [33]. Although the association of dehydration and AD is supported by these publications, the specific pathogenic nature of this association remains obscure [33, 129, 147].

3.6.2 Liver function

The rules found suggest that having high serum levels of AST (aspartate aminotransferase), as well as having low or high serum levels of ALT (alanine aminotransferase) may predispose to an impaired cognition characterized by low mini mental state examination (MMSE) scores. It should be noted that low ALT was much more rare

in the CAMD database than high ALT, so the negative effect should be attributed mainly to high ALT. However, serum ALP (alkaline phosphatase) levels seem to have a controversial effect on mental status.

AST, ALT and ALP levels derive from the liver. Elevated ALP might indicate bile duct obstruction. AST or ALT may elevate in a number of cases of liver injury or damage, spreading from acute or chronic viral infections to alcohol induced or non-alcoholic steatohepatitis. It is interesting to note that elevated serum levels of AST (more than those of ALT) have been associated with impaired mental status. Although mild elevations in serum levels of AST and ALT are nonspecific to the etiology of liver injury, certain alteration patterns in these parameters may reflect the nature of the hepatic disease. For instance, the value of the AST/ALT ratio—also known as the De Ritis ratio—is approximately 0.8 in normal subjects, a ratio exceeding 2.00 being suggestive to alcoholic hepatitis.

Therefore we scanned the subjects with high AST values for higher than 2 AST/ALT ratio: we have only found 10 subjects satisfying both conditions. In addition, only 2 rules had AST/ALT on the left-hand side. Consequently, we may assume that high serum AST in the study subjects is not typically accompanied with high De Ritis ratio (i.e. probable alcoholic hepatitis).

The association of impaired liver function with mental decline can be illuminated in two perspectives. On one hand, impaired liver function might be insufficient to prevent the brain from the effects of certain neurotoxins, e.g. ammonia. This happens in the case of hepatic encephalopathy (HE), when severe liver damage resulting in acute liver insufficiency cannot detoxificate ammonia and other neurotoxins. On the other hand, the association of elevated AST/ALT ratio with impaired mental status proposes that another obscure element (e.g. chronic alcohol consumption) might be the factor responsible for both cognitive and metabolic damages.

Our results raise the possibility of a pathogenetic linkage between liver function and mental status in patients with AD. Such linkage has also been proposed by other studies [15, 131]. One study concludes that peripheral reduction of β-amyloid is sufficient to reduce brain β-amyloid and proposes that β-amyloids, which are of

major pathogenic importance in AD may originate from the liver [131]. Another research found that deficient liver production of docosahexaenoic acid (a neuroprotective fatty acid) correlates with impaired cognitive status in AD patients [15].

To rule out the possibility when the elevated AST level is due to some medications taken, we compiled a detailed Table_S3 (in the supporting on-line material) containing the number of subjects taking certain drugs, and the number of drug-takers with high AST. The data shows that, for example, 1929 subjects took Donepezil, while among the Donepezil-takers, only 415 have had high AST levels.

3.6.3 Vitamin B12

Our results were able to present the beneficial impact of high levels of vitamin B12, also known as cobalamin, on cognition. Along with folate, vitamin B12 has an important role in the maintenance of genome integrity [52]. Although previous publications found association of low serum levels of vitamin B12 and AD [100, 102], a recent systemic review on vitamin B12 status and cognitive impairment fails to declare a clear association between vitamin B12 status and dementia [110]. However, this review also found that studies using newer and more specific biomarkers of vitamin B12 status such as methylmalonic acid and holotranscobalamin were able to draw an association between mental decline and poor vitamin B12 status [110].

Although clinically vitamin B12 deficiency may result in macrocytic anaemia, in the case of AD patients the occurrence of macrocytic anaemia is rare and the neurological and hematological features are unrelated [102].

3.6.4 Blood cholesterol

The positive or negative effects of high cholesterol values to Alzheimer's disease and cognition is a controversial issue. Some studies (e.g., [72, 152, 167]) show negative effects of high cholesterol value for cognition, while other studies ([104, 116, 117]) prove the positive effects for cognition.

Our data supports both conclusions in a sense. That is, low, low-normal and high cholesterol levels are all associated with impaired mental status, but with a different extent (scores -21, -13 and -42, respectively). See Table 9 for a selection of cholesterol-related rules from the larger Table S1 in the on-line supporting material.

It is worth to note that, by Table 9, elevated, low or low-normal cholesterol levels do not necessarily mean a higher likelihood of impaired cognition by themselves, but only combined with high sodium.

A most recent study [112] shows that the neuronal expression of amyloid precursor protein APP controls the cholesterol 24-hydroxylase mRNA levels and decreases cholesterol turnover; therefore in certain setups, the presence of amyloid precursor proteins imply lowered cholesterol levels.

3.6.5 Hematological parameters

Additional interesting rules were detected regarding hematological parameters. In particular, independently from each other, high values of mean corpuscular hemoglobin (MCH), low values of mean corpuscular hemoglobin concentration (MCHC), and low values of mean corpuscular volume (MCV) were also associated with high MMSE scores. Although high values of MCH and low values of MCHC are present in the case of macrocyctic anaemia (with the addition of high levels of mean corpuscular volume, low levels of hemoglobin and hematocrit), such solely associations should not be discussed, as they may be coincidental.

Among the rules with lift values exceeding 2.00, other parameters of hematological status (such as level of hemoglobin, red blood cell number, white blood cell number) were also present. Monocyte and eosinophil levels also appear on the left hand side of many rules with high lift. These premises appear in combinations with various other (mostly non-hematological) premises.

3.7 Conclusion

A 6000-patient, high-quality database was analyzed with original methods for biomarkers of dementia. We have found some novel and also some already well established relations connected to good or bad cognition in a 6000 patient database. The already established findings prove the validity of our data mining approach, and the new findings, related to MCH, ALP and AST levels prove its power. Some more controversial biomarkers, including cholesterol level, were also re-discovered, and we found that the high cholesterol levels seem to be beneficial only with combination with old age. The algorithm and program we developed may prove useful for others as well, for analyzing other datasets using combinatorial association rule mining.

4 The metagenomic telescope

4.1 Introduction

Next generation sequencing technologies made possible the discovery of numerous new microbe species in diverse environmental samples. Some of the new species contain genes never encountered before. Some of these genes encode proteins with novel functions, and some of these genes encode proteins that perform some well-known function in a novel way.

In this section we describe a tool, named the Metagenomic Telescope, which applies basic methods of artificial intelligence and seems to be capable of identifying new protein functions even in well-studied model organisms. As a proof-of-principle demonstration of the Metagenomic Telescope, we considered DNA repair enzymes. First we identified proteins in DNA repair in well–known organisms (i.e., proteins in base excision repair, nucleotide excision repair, mismatch repair and DNA break repair); next we applied multiple alignments and built hidden Markov profiles for each protein separately, across well–researched organisms; next, using public depositories of metagenomes, originating from extreme environments, we identified DNA repair genes in the samples. While the phylogenetic classification of the metagenomic samples are not typically available, we hypothesized that some very special DNA repair strategies need to be applied in bacteria and Archaea living in those extreme circumstances. It is a difficult task to evaluate the results obtained from mostly unknown species; therefore we applied again the hidden Markov profiling: for the identified DNA repair genes in the extreme metagenomes, we prepared new hidden Markov profiles (for each genes separately, subsequent to a cluster analysis); and we searched for similarities to those profiles in model organisms. We have found well known DNA repair proteins, lots of proteins with unknown functions, and also proteins with known, but different functions in the model organisms.

The vast field of computer science, called artificial intelligence (AI), offers great methods for distilling relevant information from large sets of data. Metagenomic

databases have been increasingly used in the recent years to investigate the bacterial composition of samples taken from a variety of environments. To analyze and compare different genomic data, Hidden Markov Models [23] provide a useful methodology.

A Hidden Markov Model, applied to protein sequences, is basically a random amino acid sequence generator with multiple internal states, two of which are distinguished as START and STOP states. The generator starts from the START state. Until it arrives to the STOP state, it repeats the following two steps:

- it outputs a random amino acid, then
- it moves to a random[1] new state.

The role of the multiple internal states is that the probability distribution of the output amino acid and the distribution of the new state both depend on the current state. The model is named "hidden" because the internal states cannot be unambiguously determined by observing the output sequence.

HMMs are particularly useful because they can be trained by a set of input sequences to output similar sequences: if we have proteins of related functions, then we can build a Hidden Markov Model which will generate random amino acid sequences as output, similar to the ones used in training. It is even a more useful property of HMMs that if we take any amino acid sequence, denoted by w, our model can easily tell us the probability of generating exactly that sequence w as an output.

Consequently, if we have a HMM trained on a certain set of proteins, then the same HMM can assign higher scores (i.e., probabilities) to proteins, *similar* to the training set, and lower scores (i.e., probabilities) to proteins, *dissimilar* to the training set. Note that this scoring is usually not homogeneous as in the case of BLAST [9] and its clones: in HMM models conservative subsequences are differentiated from those appearing in variable regions.

[1]Here the word "random" does not imply uniform distribution.

In the present work, we have applied HMM in a novel way to suggest and possibly discover still unknown protein functions in several well-studied model organisms. Starting from sequence alignments for proteins involved in DNA damage repair, we created Hidden Markov Models and used these models to search for similar genes in the metagenomic samples from different environments. Combining the original HMM with the genes found in the metagenomes, we created a second, more trained HMM that we used to interrogate proteomes of higher order model organisms. This search (termed as "the Metagenomic telescope" in the present study) generated numerous novel hits in the higher order organisms, containing proteins previously not yet described as closely similar to the DNA damage repair proteins. These results indicate the Metagenomic Telescope may be a powerful method for the identification of novel proteins in higher order model organisms.

4.2 Methods

First, we took some known *E. coli* and Archaean occurrences of a specific enzyme as listed in Table 1. We aligned these similar proteins using Clustal Omega [125]. The aligned sequences were then used to train a HMM with the `hmmbuild` utility of the HMMER3 package [50]. We call the resulting model the "original HMM".

This "original HMM" was used twice: once in the direct projection to the model organisms, and second time for Projection 1 in the Telescope.

In the original projection, similarity scores are assigned to the protein sequences of the model organisms: the output of this single projection is the set of the highest scored proteins found in the proteomes of the model organisms (this step is visualized on the upper panel of Figure 4).

For the application in the Telescope, we first extracted open reading frames from the metagenomes with the `getorf` application of EMBOSS [118], then run the `hmmsearch` utility of HMMER3 [50] on the "original HMM" and the database of amino acid sequences extracted from each metagenome. The result of this search consist of hits in the metagenome can be referred to as "metagenome matches".

Three extreme metagenomes in the present study were accessed through the CAMERA portal [124]:

Richmond Mine in Iron Mountain: CAMERA accession code: CAM_PROJ_AcidMine. The Iron Mountain, California mine was closed in the sixties, and the large pyrite deposits exposed to atmospheric oxygen and moisture produce one of the most acidic mine drainage on Earth [16]. The metagenome consists of data gained by sequencing samples from the thick, pink biofilm in this acidic and hot (42 °C) environment, containing iron-oxidizing bacteria and other species.

Yellowstone Bison hot spring: CAMERA accession code: CAM_PROJ_BisonMetagenome. The Bison Pool environment is an alkaline hot spring in the Sentinel Meadow of Yellowstone National Park, in Wyoming. The samples were collected from sites with water temperature of 92 °C through 56 °C [47, 70, 132].

Phosphorus removing (EBPR) sludge community: CAMERA accession code: CAM_PROJ_EBPRSludge. The samples are originated from an enhanced biological phosphorus removal (EBPR) sludge community from Thornside Sewage Treatment Plant in Brisbane, Queensland, Australia.

The metagenome matches were aligned and clustered using the OPTICS method [12]. The clusters were then used as inputs of `hmmbuild` [50], which yielded the "new HMMs". In other words, these models have been built on possible unknown DNA repair enzymes found in the metagenome. We then performed the final step in the process pipeline, i.e., testing both the original and the new HMM's for the proteomes of higher level organisms. As visualized on Figure 4, we compared the results of the projection on the upper panel and the projections of the lower panel. These organisms included *Arabidopsis thaliana, C. elegans* and *E. coli* as well as mouse, rat, human, and other model species. The flowchart of the application of the Metagenomic Telescope is given on Figure 3.

Our goal was to examine whether the possible new DNA repair enzymes found in the metagenomes could be used for finding new DNA repair enzymes in the model organisms as well. This included comparison of the results of the searches performed

with respectively the original and the new models.

4.3 Design of the Metagenomic Telescope

The optical (refractive) telescope applies two projections: the first projection is done by the objective lens, the second by another lens called "the eyepiece": through the eyepiece one can see the enlarged image, generated by the objective.

Our Metagenomic Telescope also consists of two projections, each are performed by applying HMMs. The key point is making use of *metagenomes* in projections: *first* we project *to* metagenomes, then we project *from* metagenomes. The lower panel of Figure 4 describes these two projections, and compares these to a single HMM projection on the upper panel of Figure 4.

The starting point is a set of proteins of similar function or structure, taken from well–annotated organisms. This set is the teaching set of the first HMM in the Metagenomic Telescope and the single HMM the original projection.

In the original HMM or the original projection (upper panel of Figure 4), we use the HMM constructed in this step for finding similar protein sequences in model organisms: this is the only projection we use here. Using that HMM, similarity scores are assigned to the protein sequences of the model organisms. The output of this projection is the set of the highest scored proteins found in the proteomes of the model organisms.

In contrast, in the Metagenomic Telescope (lower panel of Figure 4), we apply two projections:

Projection 1 in the Telescope: Here we use the same HMM as in the original projection, but now we search for high-scored protein sequences from the metagenomes instead of proteins from the model organisms.

Projection 2 in the Telescope: The starting point is the highest scored proteins from the metagenome. After a suitable clustering, new – second, third, ... – HMMs are built: the teaching sets consist of these high scored proteins. Next, the proteomes

of some model organisms are considered, and by the new HMMs, similarity scores are assigned to the protein sequences of the model organisms. The output of the second projection is the set of the highest scored proteins found in the proteomes of the model organisms.

The two concepts: the original HMM projection and the Metagenomic Telescope, are detailed on Figure 4.

We believe that our telescope would help to view and annotate more clearly the functions of proteins in model organisms, since the diversity of well-chosen metagenomes would help to assign new, still unknown functions to a number of proteins.

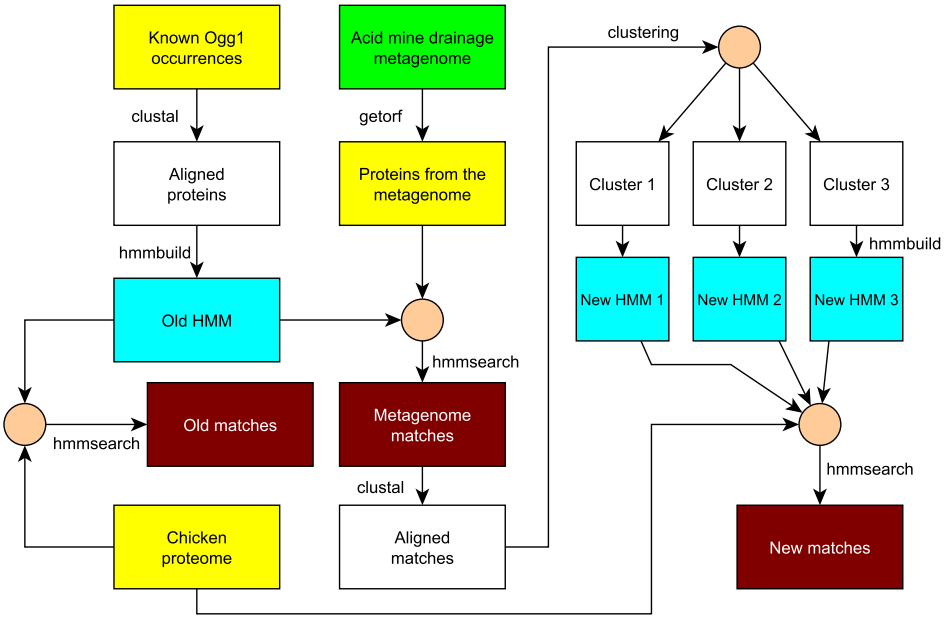

Figure 3: The flowchart of the Metagenomic Telescope applied to DNA repair enzymes.

4.4 Proof of concept: DNA repair enzymes

As a proof of concept, we apply the Metagenomic Telescope to DNA repair enzymes as the starting set of proteins, and metagenomes, found in extreme environments (acid mine leakage, a Yellowstone hot spring and a phosphorus removing sludge com-

munity), see Table 1. Since organisms living in extreme environments are supposedly suffer more frequent DNA damage than organisms in ambient conditions, we assume that their DNA repair mechanisms are more efficient than that of other organisms. Therefore one can hope to find new, more efficient DNA-repair enzymes in these extreme metagenomes. Certainly, there is a remarkable scientific interest in finding novel, more efficient enzymes in exotic species of the metagenomes mentioned. However, there is a much stronger interest in finding new functions for already known enzymes and functions for proteins with unknown role in important model organisms, including *Homo sapiens*. Therefore we perform a second projection from the DNA-repair enzymes to several model organisms.

4.4.1 HMM projections of single-domain proteins

Among the protein families involved in DNA damage recognition and repair selected for this present study, the trimeric dUTPase family constitutes a well defined protein fold which can be also found in the family of prokaryotic dCTP deaminases. In eukaryotes, however, to our knowledge no other proteins have yet been described that show this peculiar fold. Also, eukaryotic dUTPases are described as monogenic in the model eukaryotic organisms studied to date. In accordance with this strong "stand-alone" character of this protein family, HMM searches found the dUTPase orthologuos sequences, however, no novel protein could be found among the original hits. Still, among the telescopic hits, we found one novel hit in the mouse proteome (UniProt accession number Q3TL09). Although on the sequence level it showed rather low similarity to the authentic dUTPase sequence (identitity 9%, similarity 23%), still the alignment shows that out of the five characteristic dUTPase motifs, four can be identified in the sequence of this protein (Figure 5A). The actual functional relevance of this protein to dUTPases needs further experimental studies out of the scope of the present work. It was also of interest to investigate if the 3D structure of this protein may be similar to the dUTPase fold. For such investigations, first we run the SwissModel software [13,93] by nominating the human dUTPase 3D structure (PDB ID 3EHW) [106, 148, 149] as the template. Results showed that the

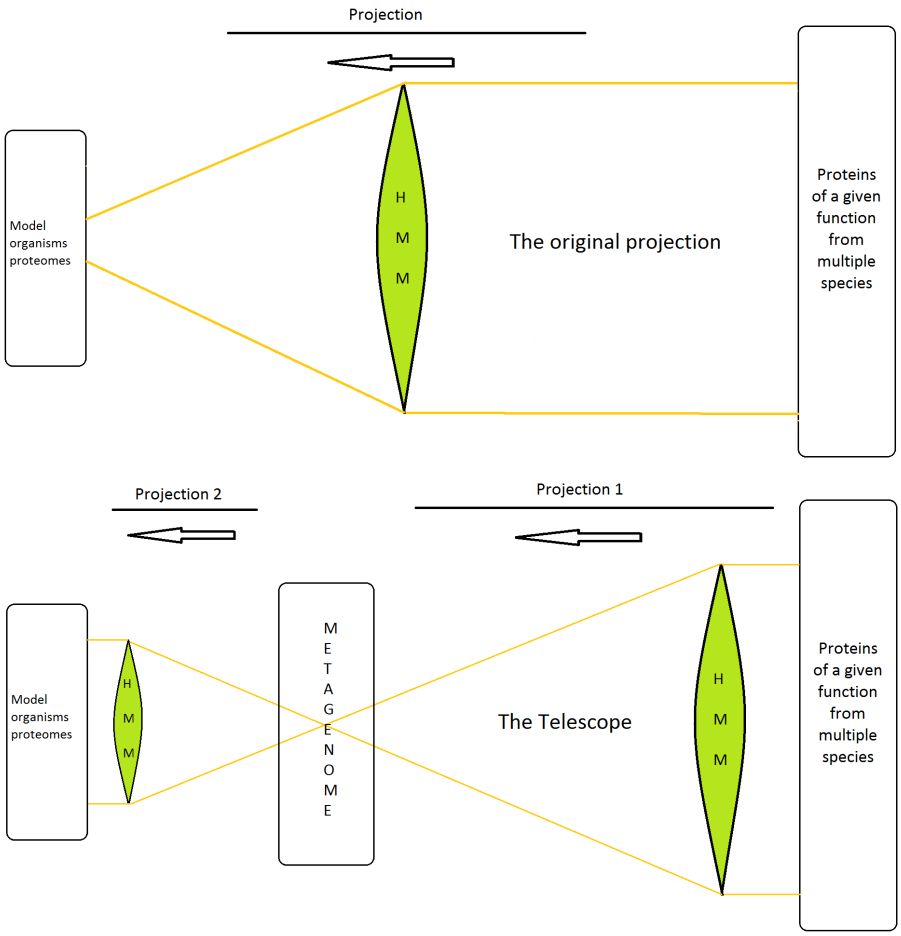

Figure 4: The original projection (upper panel) compared to the scheme of the Metagenomic Telescope (lower panel). Projection 1 discovers genes or proteins in the metagenome that probably have similar function as the well-known starting proteins in front of the objective lens on the right hand side.

dUTPase fold can be adopted by this protein (Figure 5B), however, the strength of this conclusion is somewhat weakened by the fact that the template was pre-defined and could strongly perturb the results.

Hence, we next used the MUSTER software [162] without any pre-defined template. This recently described software is based on an integrated use of protein profiling information and tries to fit a 3D structure from the Protein Data Bank on the sequence submitted. Results of the MUSTER-modeling showed that three slightly different 3D models could be created, and very interestingly, all of these

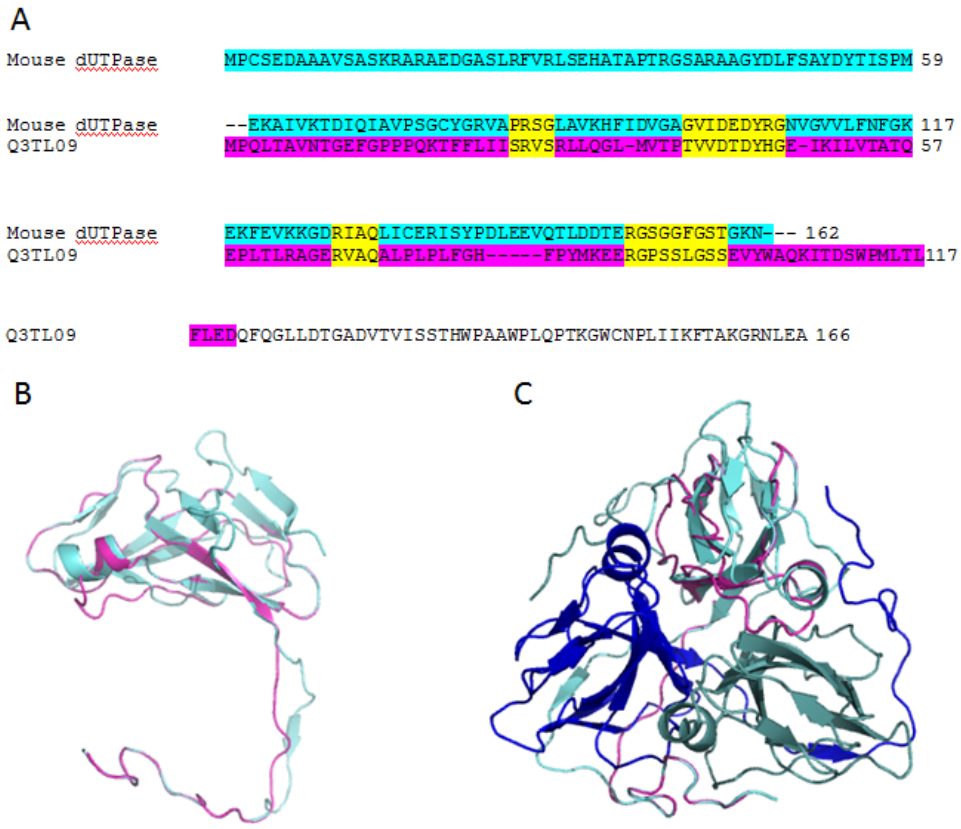

Figure 5: Identification of a novel dUTPase-like protein in the mouse proteome using the Metagenomic Telescope. Panel A shows an alignment (created by ClustalW) between the mouse dUTPase sequence (cyan) and the novel hit associated with the Uniprot ID Q3TL09 (purple color indicates the part of this latter sequence that could be modeled in 3D using SwissModel or MUSTER). The conserved dUTPase motifs are shown in yellow. Panel B illustrates the structural alignment between human dUTPase (cyan) and the modeled Q3TL09 structure (purple) (at the subunit level). Panel C shows one of the models for Q3TL09 created by MUSTER software (purple), in this case the trimeric structure characteristic of dUTPases is shown (monomers are in shades of blue: cyan, royal blue and grayish blue). Protein structural models are shown in ribbon diagrams (PyMol).

used a dUTPase structure as the best-fitting model (Figure 5B).

We conclude that for the dUTPase searches, the use of the telescopic HMM resulted in a promising finding. The newly found mouse protein, although with a very low level of sequence identity, may adopt the 3D structure of the antiparallel beta-sheeted jelly roll dUTPase-fold.

HMM models were also created for the numerous DNA-glycosylase families (listed

in Table 1) that belong to either the alpha/beta superfamily of uracil-DNA glycosylases (UNG, TDG) or to the helix-turn-helix (HTH) superfamily of DNA glycosylases (NTH, NEI, OGG) [87]. These proteins, similarly to dUTPases, are also single domain proteins, with some N- or C-terminal extensions in several eukaryotic organisms. In several cases, eukaryotes encode different isoforms of DNA-glycosylases, dedicated to the different cellular compartments (nuclear vs. cytoplasmic). We found that while the original hits usually included the orthologues and their isoforms, the telescopic hits also included hits from the whole superfamily. For example, starting with the uracil-DNS glycosylase UNG, original hits showed the orthologous nuclear and mitochondrial isoforms of UNG, while telescopic hits included the closely related thymine-DNA glycosylases as well as SMUGs. Similarly, starting from any of the HTH superfamily DNA glycosylases, original hits were rather restricted to the different isoforms of the same proteins, while telescopic hits included proteins of the whole HTH superfamily.

Protein families	**Archea proteomes involved in building the "original HMM"**	**Metagenomes involved in building the "telescopic HMM"**	**Eukaryotic proteomes screened by HMM models**
dUTPase	*Haloferax volcanii*	Iron Mountain acid sludge	*Saccharomyces cerevisae*
uracil-DNA glycosylase (UNG)	*Halobacterium salinarum*	Yellowstone Bison hot spring	*Arabidopsis thaliana*
thymine-DNA glycosylase (TDG)	*Methanobacterium thermoautrophicum*	Phosphorus removing sludge community	*Caenorhabditis elegans*
Archeal UDG	*Methanococcus maripalidus*		*Drosophila melanogaster*
NTHL1	*Methanococcus janaschii*		*Danio rerio*
OGG1	*Methanosarcina acetivorans*		*Gallus gallus*
	Thermococcus kodakaraensis		*Bos taurus*
Rad50	*Archeoglobus fulgidus*		*Canis lupus*
Mre11	*Aeropyrum pernix*		*Mus musculus*
	Pyrococcus furiosus		*Sus scrofa*
	Pyrococcus abyssi		*Rattus norvegicus*
	Pyrococcus horikoshii		*Homo sapiens*
	Sulfolobus acidocaldarius		
	Sulfolobus islandicus		
	Sulfolobus solfataricus		

Table 1: Protein families, metagenomes and proteomes used in the present study.

4.4.2 HMM projections of multiple domain proteins

The Mre11 and Rad50 proteins play important roles in the repair of double-strand-DNA breaks. These two proteins are essential in both major pathways of double-stranded DNA break repair, in homologous recombination repair, as well as in non-homologous end-joining. Both Rad50 and MRE11 are multidomain proteins (c.f., Figure 6). Rad50 has an ATPase globular domain and a highly lengthened coiled-coil domain connected together with a Zn-hook, whereas Mre11 contains a phosphodiesterase core domain and several DNA-binding recognition loops.

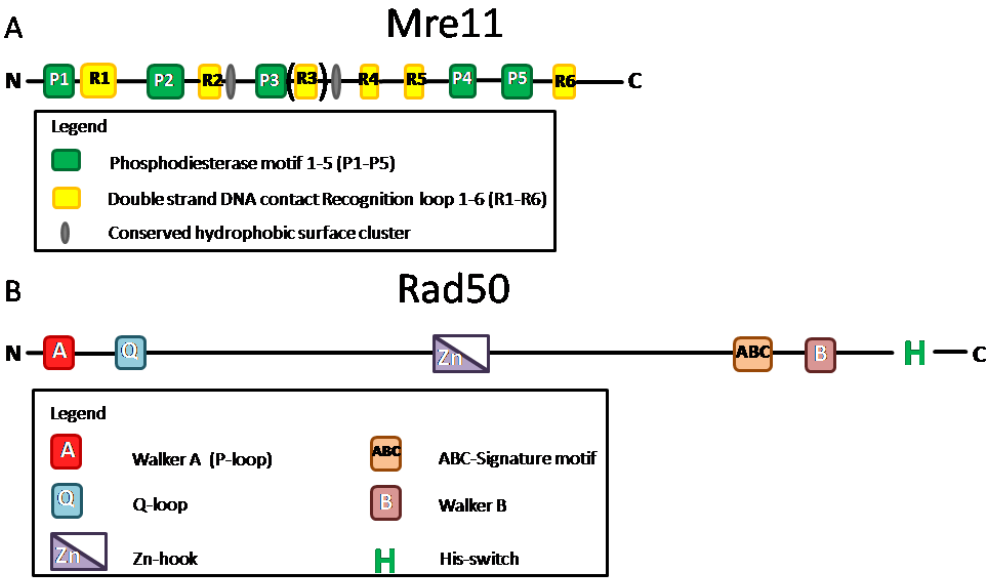

Figure 6: Schematic representation of Mre11 (Panel A) and Rad50 (Panel B) domains. (A) Mre11 has five phosphodiesterase motifs (green), 6 dsDNA recognition loops (yellow) and hydrophobic surface clusters (gray) (B) Rad50 has a bipartite ATPase domain: Walker A (red), Walker B (pale red), Q-loop (light blue), ABC-Signature motif (orange), histidine switch (green H) and a Zinc-hook (purple).

Rad50 and Mre11 usually form a heterotetramer and this assembly is termed as the MRN complex. The MRN complex is crucial to (i) bridging DNA over short and long distances, a(ii) DNA binding and processing and (iii) activation of double strand break response and checkpoint signaling pathways [154]. Both Mre11 and

Rad50 need metal cofactors: manganese and magnesium, respectively [77]. Both of them can bind DNA. The dimerization of Rad50, which belongs to the ABC-ATPase family [78], is ATP dependent [78]. Rad50 has a conserved "signature motif" that is needed for binding the γ-phosphate of ATP and is characteristic for ABC-ATPases [78]. The "signature motif" has a key role in Rad50 dimer assembly [78]. The Walker A motif binds ATP, while the Walker B motif hydrolyses it [77]. The Walker A motif (also called P loop or phosphate-binding loop) also forms the nucleotide binding site [77]. The D loop, a part of Walker B, binds one active magnesium ion and assists in dimerization [78]. The Mre11 binding site is on the coiled-coil region near the ABC domain [77].

Mre11 has five conserved phosphodiesterase motifs [77]. Conserved hydrophobic surface clusters are likely involved in macromolecular interaction sites [77]. The six DNA recognition loops (R1-R6) constitute a continuous DNA interaction surface [155]. All core DNA recognition loops are conserved in *S. pombe, S. cerevisiae* and *Xenopus*, except recognition loop 3 (R3) [155]. Rad50 and Mre11 homologues in *Escherichia coli* are termed as SbcC and SbcD, respectively [43, 45].

The results of the application of the Metagenomic Telescope on these protein families are summarized in Figures 7 and 8 (for Mre11 and Rad50, respectively). In both figures, one panel (Figure 7A and 8A) shows the actual number of hits found in the original as well as in the telescopic projections in the model eukaryotic organisms. This representation provides a rather straightforward measure of the strength of the telescopic projection over the original ones. In some cases, the number of hits is just 1 (e.g., in the case for the original hits of Mre11 in several model organisms). In these cases, the hit was actually the *bona fide* Mre11 homologue in the given organism, and no additional "similar" proteins can be found. However, in the majority of cases, the number of hits is more than 1, and in these cases, in addition to the *bona fide* homologue that was always within the hits, additional proteins were also identified by the HMM projections.

The fact that the *bona fide* homologue is always identified shows that the HMM projections are reliable. Nevertheless, these are the additional hits that may contain

novel properties. It is easy to see for both Mre11 and Rad50 that the number of hits using the telescopic projections are never smaller than for the original projections, on the contrary, these hits are quite frequently significantly more numerous. The additional hits, identified only in the telescopic projections are termed "new telescopic hits" on the respective panels in Figures 7A and 8A.

To analyze the putative biological functions of the original and the new telescopic hits, in each cases we relied on the genome ontology classification categories and listed the different genome ontology definitions for each hit. The biological functions (genome ontology categories) found to be associated with most of the original hits are rather straightforward to assess. Accordingly, for both the Mre11 and Rad50 families, we find that the functions listed (metal binding, DNA binding, DNA repair, etc) are already known to be associated with the Mre11 and Rad50 families.

Next, we compared the original and telescopic hits and found that the list of these properties is significantly enriched in the telescopic hits. Therefore, not only the number of hits was higher after using the telescopic HMMs, but also these hits were associated with additional functional properties (Figures 7B and 8B).

In order to evaluate the power of the Metagenomic telescope method, we need to consider those genome ontology terms that show up only in the new telescopic hits. For the Mre11 family, such terms are the calcineurin-like phosphoesterase (CPPED1) family, the metallophosphoesterase family and the acid phosphatase biological function. While the latter two may be explained on the well-known characteristics of the Mre11 enzymatic action, the connection to the calcineurin-like phosphoesterase family seems to be novel. In this case, at least to our knowledge, the potentially similar characteristics of Mre11 and calcineurin-like phosphoesterases have not yet been addressed before. In the case of the hits within the Rad50 family, the novel hits using the telescopic projections are even more evident. Perhaps the most intriguing result from these projections concerns the numerous occurrence of the "transcription regulation" and "transcription factor" genome ontology classes, which are evidently linked. Based on these findings, we suggest that Rad50-like proteins may also be involved not just in interacting with DNA but also interacting with the transcrip-

tion process. It is known that e.g., DNA damage and repair occurs with higher frequency on transcriptionally active genomic segments since these are more accessible. Our present results may suggest that, in addition to the less physical barrier in the actively transcribed genomic regions, Rad50-like proteins may also be involved in interacting with the transcription machinery in a more direct manner.

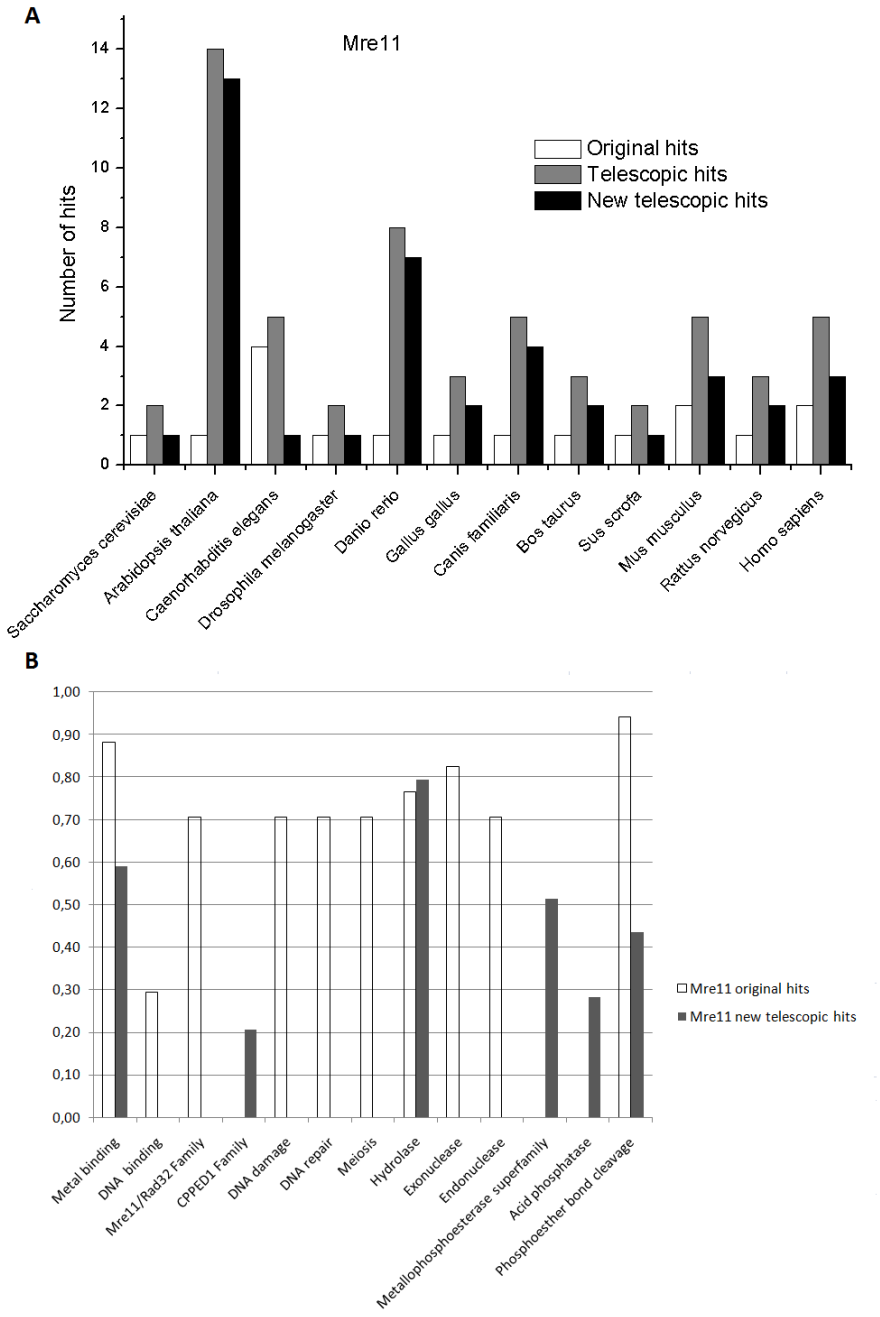

Figure 7: Original and telescopic hits for the Mre11 family. Panel A. Number of hits identified in the various eukaryotic model organisms after the original and the telescopic projections. Panel B. Distribution of genome ontology terms within the different hits. Note that new genome ontology classes can be observed in the telescopic hits.

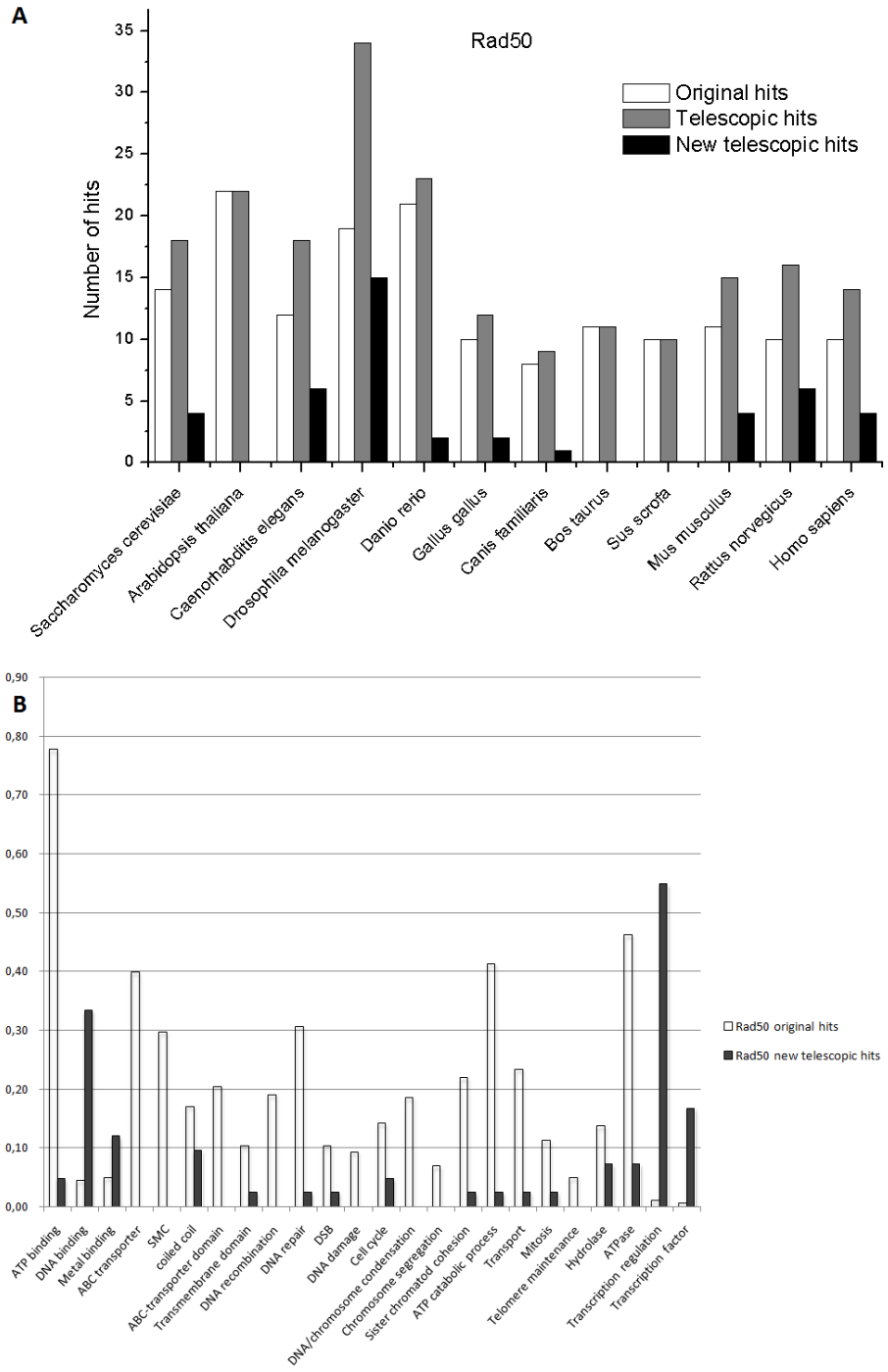

Figure 8: Original and telescopic hits for the Rad50 family. Panel A. Number of hits identified in the various eukaryotic model organisms after the original and the telescopic projections. Panel B. Distribution of genome ontology terms within the different hits. Note that new genome ontology classes can be observed in the telescopic hits.

5 Nucleotide 9-mers and diabetes

5.1 Introduction

In the present section we describe nucleotide 9-mers with significantly differing frequencies in diabetic and healthy intestinal flora. To our knowledge, it is the first time such short DNA fragments have been associated with T2D. The automated, quantitative analysis of the frequencies of short nucleotide sequences seems to be more feasible than accurate phylogenetic and functional analysis, and thus it might be a promising direction of diagnostic research.

Discoveries of new biomarkers for frequently occurring diseases are of special importance in today's medicine. While fully developed type II diabetes (T2D) can be detected easily, the early identification of high risk individuals is an area of interest in T2D, too. Metagenomic analysis of the human bacterial flora has shown subtle changes in diabetic patients, but no specific microbes are known to cause or promote the disease. Moderate changes were also detected in the microbial gene composition of the metagenomes of diabetic patients, but again, no specific gene was found that is present in disease-related and missing in healthy metagenome. However, these fine differences in microbial taxon- and gene composition are difficult to apply as quantitative biomarkers for diagnosing or predicting type II diabetes.

Metagenomics [40] is rapidly gaining importance in clinical research [10,27,31,34, 98,107,115,164], environmental studies [54,111,165] and biotechnology [53,123,128]. Numerous complex and reliable methods have been published for the phylogenetic identification of non-cloned short DNA reads from environmental or clinical samples, for example, the similarity-based methods MEGAN [80–82] and MG-RAST [62,153], the marker gene based phylogenetic analyzer AMPHORA [160] and its more user friendly versions, AMPHORA2 [161] and AmphoraNet [91, 92].

These methods use multi-phase, complex approaches to retrieve phylogenetic information from the short read datasets, applying reference database operations in the process.

Surprisingly, it was shown that simple frequency counting of nucleotides or short nucleotide sequences in the metagenomic samples may also imply phylogenetic information.

It has been widely known for a long time that genomic AT/GC ratio is distributed in a wide range in bacterial species, and can be characteristic to some of them [25, 79, 130]. The ratio is shown to be influenced by numerous environmental and metabolic factors [159] and also carries phylogenetic information.

The article [90] reports differences in di- and tetranucleotide frequencies among numerous bacterial species, and examines the possible application of these signatures in molecular phylogeny.

Tetranucleotide sequence frequencies were applied in supervised and unsupervised phylogenetic classification, or "binning" in [122].

The work [95] applies conserved gene fragments, each encoding several dozens of amino acids, identified from the Pfam database [126]. The fragments are called "environmental gene tags", and are used successfully for phylogenetic binning in [95].

The study of [115] investigated the differences in gut metagenomes of diabetic and healthy subjects. The metagenomes were *de novo* assembled, and the bacterial genes were mapped to a metagenomic gene catalog. Genes related to oxidative stress response were found more abundant in the samples originating from diabetic subjects. Additionally, moderate changes in intestinal bacterial composition were detected, but no specific microbes were associated with the metagenomes of type II diabetes (T2D) patients.

After a very complex selection and filtering process, genome-specific nucleotide markers of length 50 were identified in [146]. The markers were applied for strain/species identification, and also as markers for microbial species that might play a role in T2D and obesity in the data set of [115].

Here we describe a very simple and straightforward approach for finding short nucleotide sequences whose frequencies significantly differ in T2D and healthy metagenomes of the dataset of [115]. We identify several nucleotide 9-mers that may serve as quan-

titative biomarkers of the pre-diabetic state in the future. To our knowledge, such short sequences have never been found to characterize T2D or any other disease.

We need to clarify that we do not state that the identified 9-mers will generally be applicable as biomarkers for diabetes for all human populations. We believe that "enterotype-specific" [14] quantitative biomarkers could be found for each enterotypes by exhaustive searches described in the Methods section, and those enterotype-specific biomarkers could serve as predictors of type 2 diabetes mellitus.

5.2 Methods

Our data source was the set of metagenomes of 345 Chinese subjects, collected by Qin et al. [115] and deposited in the Sequence Read Archive (http://www.ncbi.nlm.nih.gov/sra) under accession numbers SRA045646 (145 subjects) and SRA050230 (225 subjects). The assembled data was downloaded from the GigaScience database, GigaDB at http://dx.doi.org/10.5524/100036.

We considered all the possible DNA sequences of length at most 9 (this means over 300,000 possible sequences). For each sequence, we counted the number of exact matches in each raw metagenome. Our aim was to determine whether there are any short DNA fragments whose frequencies differ for diabetic/non-diabetic, lean/obese or female/male individuals. We had to draw the line at 9 nucleotides, because calculating the frequency of longer sequences is computationally more expensive, and the chance of a false positive greatly increases when testing a large number of sequences.

We first defined the frequency of a short DNA fragment for a given metagenome as the number of occurrences (exact matches), divided by the total size, measured in base-pairs (bp), of the metagenome. Additionally – to account for minor mutations – we also included those sequences in the counting process that differed by only one nucleotide, but these were considered with half a weight. So, for example, the final *frequency* of the sequence AAA included not only how many times the sequence AAA occurs in a specific metagenome, but also how many times AAG, CAA, ATA,

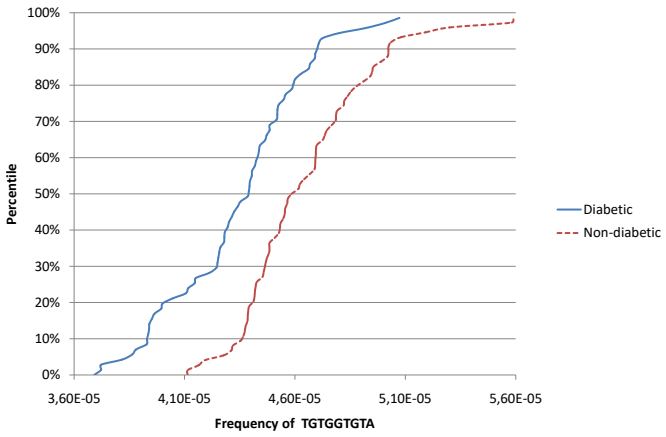

Figure 9: Empirical cumulative distribution function of the frequency of TGTGGTGTA (solid: diabetic, dashed: non-diabetic). For every y, the curves show the frequency of TGTGGTGTA at the yth percentile among the diabetic (solid line) and non-diabetic (dashed line) fraction of metagenomes. For example, for $x = 0.000045$, the frequency of TGTGGTGTA is less than x in 70% of the diabetic samples, but in only 38% of the non-diabetic samples. Further empirical cumulative distribution functions are in Figure 10.

... occur in that metagenome, except that the number of occurrences for these related DNA fragments was divided by two.

Let ℓ_M denote the length in base-pairs (bp) of a metagenome M. Let $d(s,t)$ be the number of mismatches between the two sequences of same length, s and t (also called the Hamming distance). Let $k_M(s)$ denote the number of exact matches of sequence s in metagenome M. Then $f_M(s)$ (the frequency of sequence s with respect to metagenome M) is defined by the formula

$$f_M(s) = \frac{1}{\ell_M}\left(k_M(s) + \frac{1}{2}\sum_{d(s,t)=1} k_M(t)\right).$$

This approach (counting some non-exact matches as well, but with the half the weight) yielded statistically better results when compared to the original, stricter counting process, which only allowed exact matches.

We developed C++ programs for counting the fragments and analyzing the results. Several partitions on the set of subjects were analyzed, by dividing them into two groups by different attributes: diabetic/non-diabetic, lean/obese and female/-male. Our aim was to look for short DNA sequences whose mean frequency differs for the two groups.

To achieve this, first we calculated $f_M(s)$ for each raw/assembled metagenome M and each short DNA sequence s of length $\ell_s \leq 9$. Then, for each s we calculated a p-value using Welch's t-test, which showed whether the frequency s is the same in the two groups (i.e., p is large) or differs significantly (i.e., $p \leq 0.05$).

Since this was done for each short DNA fragment, the number of total statistical tests done for a given division of subjects was equal to the number of possible s DNA sequences of length at most 9. As this is more than 300,000, there was a high probability that one of the tests would yield a very low p-value but the large measured difference of means would be in fact due to mere chance. On the other hand, applying an FDR or FWER control procedure directly (instead of the two-step method) would have resulted in multiplying the p-values with a very large number, which would have yielded almost no significant results.

Therefore we utilized a two-step hypothesis testing procedure. First we computed the p-values for Study 1 (with 145 subjects, SRA accession number SRA045646) only, which now became our *training set*. Then we sorted the possible s sequence candidates by p-value ascending, and chose those 20 sequences which had the lowest p-value. These were those sequences which showed promise that their frequency might differ significantly between diabetic/non-diabetic, lean/obese and female/-male individuals, depending our current partitioning of the subjects. Then we tested these selected sequences (and corresponding statistical hypotheses) on the *holdout set*, which was the collection of metagenomes from Study 2 (SRA accession number SRA050230, 225 subjects). On this set we performed only those 20 tests which qualified in the first round, which again yielded a second p-value for each of the 20 DNA sequences.

Fragment	Diabetic	Non-diabetic	p (training set)	p (holdout set)	p (corrected)	FDR
TGTGGTGTA	4.48E-05	4.71E-05	7.80E-09	0.000296	0.021151852	0.021151852
TGTGCTATC	4.35E-05	4.55E-05	1.87E-08	0.001764	0.063026802	0.063026802
TGTGGTACT	4.01E-05	4.16E-05	9.51E-10	0.001929	0.04594811	0.063026802
TGTGGTA	0.0006214	0.0006428	1.40E-08	0.001937	0.034604001	0.063026802
TGTGGTACA	4.67E-05	4.88E-05	2.97E-08	0.002098	0.029984179	0.063026802
AGTACCACA	4.10E-05	4.24E-05	2.15E-08	0.002246	0.02674947	0.063026802
CCATCTGT	0.0002318	0.0002424	2.14E-08	0.003092	0.031564443	0.063026802
TGCCACATA	5.81E-05	6.13E-05	6.42E-09	0.004678	0.041785626	0.063026802
TGTGGTATG	4.81E-05	5.04E-05	9.19E-09	0.004925	0.03910393	0.063026802
TACCACA	0.0006332	0.0006531	3.38E-08	0.004999	0.035722333	0.063026802
TGTGGAGAT	6.54E-05	6.79E-05	1.52E-08	0.008901	0.05782329	0.063026802
TGTGGTATC	5.04E-05	5.25E-05	1.49E-08	0.011902	0.070875377	0.070875377
ATGGTCTGT	5.85E-05	6.07E-05	1.29E-08	0.012383	0.068067407	0.070875377
GTACCACAT	4.18E-05	4.31E-05	1.06E-08	0.012814	0.065405364	0.070875377
CCACATACT	5.13E-05	5.35E-05	2.44E-08	0.014294	0.068095624	0.070875377
ATGTGGTAC	4.14E-05	4.27E-05	9.50E-09	0.02434	0.108706941	0.108706941
TCTCCACAT	6.97E-05	7.26E-05	1.58E-08	0.07478	0.314335349	0.314335349
ATCTCCACA	6.62E-05	6.84E-05	5.43E-09	0.078516	0.311703978	0.314335349
CTCCACATA	5.58E-05	5.75E-05	2.02E-08	0.257111	0.966993912	0.966993912
TCCACAT	0.0008132	0.0008294	1.92E-08	0.266428	0.951933372	0.966993912

Table 2. Frequencies of 7-, 8- and 9-mers in diabetic vs. non-diabetic samples with the highest significance (training set: Study 1, holdout set: Study 2). The columns of the table are: the sequence itself, the frequencies for diabetic and non-diabetic subjects, the p-value for the training and the holdout sets, the corrected p-value for the holdout set (multiplied by the factor determined by the Benjamini-Hochberg correction), and the false discovery rate for the fragments so far. Choosing an FDR of about 7% allows us to make 15 discoveries, expecting about 1 of them to be false, but the real FDR should be lower due to strong positive correlation among the tests. It is easy to recognize that TGTGGTA and TACCACA are exact complements. The complement of TCCACAT, ATGTGGA, is almost the prefix of ATGTGGTAC. 9-mer TGTGGTACT (line 3) is the exact complement of AGTACCACA (line 6). One can find further complementarities in the table. These independently found complements with very close frequencies and p-values strengthen our findings. More tables (for lean-obese and female-male distributions) are given below.

Then the Benjamini-Hochberg correction was used to determine which of the sequences had a significantly different frequency among the two groups. This correction algorithm effectively controls the false discovery rate (FDR) by calculating a q-value which takes the fact that we performed multiple (i.e., 20) statistical tests into account. Since the frequencies in the second study are independent from those in the first study, the first one is indeed a suitable training set for the model, and we can safely ignore that we performed over 300,000 statistical tests on the first study, since we use only the tests on the holdout set to make predictions. We had to utilize the version of the Benjamini-Hochberg-Yekutieli procedure which had no assumptions about dependence, since the correlation between the number of occurrences of two fragments depends on how much they overlap, and, for sequences having no overlaps, this correlation is negative. This conservative method might have resulted in calculating larger than necessary FDR values.

We have applied the raw, unassembled metagenomes from Study 1 and Study 2 to look for short marker sequences of diabetes.

Unfortunately, there was not enough information available to us to determine which subjects of Study 2 are lean/obese or female/male. Thus we had to use the available assembled metagenomes in Study 1 to look for marker fragments for sex and obesity. We partitioned the assembled metagenomes of the first study into two "random" groups: one of the groups consisted of those individuals with an odd subject ID, and the other group contained those with an even ID. One of these was the training set and the other became the holdout set, i.e. they took the role of Study 1 and Study 2 for the lean/obese and female/male classifications (Tables 3 and 4).

One sequence passed an FDR threshold of about 0.1 for the lean/obese division, and none of the short sequences had a significant difference of frequency between the two sexes.

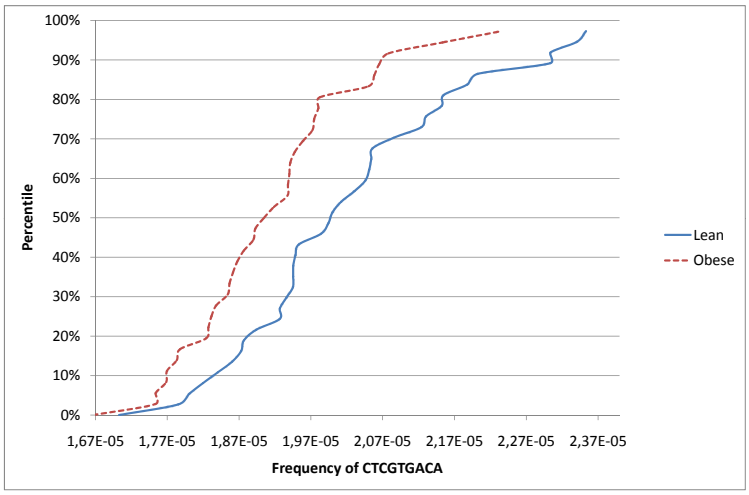

Figure 10: Empirical cumulative distribution function of the frequency of fragment CTCGTGACA (solid: lean, dashed: obese)

5.3 Discussion and results

Our results are summarized in Table 2 and Figure 9. Table 2 contains 20 7-, 8- and 9-mers of the highest statistical significance, distinguishing between the diabetic and non-diabetic metagenomes of the study [115].

Table 2 was prepared without considering complementarities between the short nucleotide sequences. Therefore, the complements found with very close frequencies and statistical parameters independently verify our results. It is easy to recognize in Table 2 that TGTGGTA and TACCACA are exact complements. The complement of TCCACAT, ATGTGGA, is almost the prefix of ATGTGGTAC. The complement of TGTGGTACT (line 3) is again the exact complement of AGTACCACA (line 6), just to mention some of the complementarities in the table.

Figure 9 gives the empirical cumulative distribution functions of the frequency of 9-mer TGTGGTGTA in the diabetic and in the non-diabetic samples. The difference between the expected values (means) of the two distribution is obvious on the figure and is quantified statistically in Table 2.

The source of the bias in short nucleotide sequence frequencies is most probably

due to the difference in the gene- and species composition of diabetic and healthy metagenomes, found in [115, 146]. These frequencies could be measured and evaluated more easily than the much more involved characteristics found in [115, 146].

5.3.1 Lean/obese and male/female classes

We also searched for short nucleotide sequences characterizing lean/obese and male/female individuals in the dataset of [115]. Only one short sequence passed even a rather large statistical significance bound of about 0.1 in the lean/obese search, and none in the male/female search (see Table 3 and 4 and Figure 10).

Fragment	Lean	Obese	p (training set)	p (holdout set)	p (corrected)	FDR
CTCGTGACA	2.00E-05	1.90E-05	0.002091	0.001443	0.103115277	0.103115277
CTCGATTGT	2.85E-05	2.73E-05	0.002945	0.004539	0.162176107	0.162176107
TGTCGACTG	2.46E-05	2.30E-05	0.0009184	0.005781	0.137701413	0.162176107
ACACTCGAG	1.13E-05	1.03E-05	0.001831	0.006911	0.123463216	0.162176107
CTCGAGTGT	1.13E-05	1.03E-05	0.002036	0.012364	0.176703713	0.176703713
TGTGACTCG	1.35E-05	1.29E-05	0.002158	0.014499	0.172680574	0.176703713
ATGTGAGGC	2.35E-05	2.26E-05	0.001805	0.016175	0.165121237	0.176703713
GTGCCTCTC	2.38E-05	2.26E-05	0.002931	0.019559	0.174708222	0.176703713
GGCTCACTC	1.82E-05	1.72E-05	0.003306	0.03181	0.252567719	0.252567719
CGAGTGAGA	1.86E-05	1.79E-05	0.003293	0.036067	0.257731026	0.257731026
CACTCGAGG	1.21E-05	1.09E-05	0.003403	0.061201	0.397578157	0.397578157
GAGTGAGCT	2.15E-05	2.06E-05	0.003223	0.062982	0.375052345	0.397578157
CTCGACTGT	2.06E-05	1.95E-05	0.003178	0.071181	0.39127078	0.397578157
CTGTCGTGT	2.72E-05	2.63E-05	0.00301	0.07767	0.396444094	0.397578157
TGTGGTTGA	5.72E-05	5.52E-05	0.002553	0.121549	0.579050998	0.579050998
CACTCGTGG	1.63E-05	1.52E-05	0.002677	0.130222	0.581595532	0.581595532
TCACCATGT	4.98E-05	4.83E-05	0.003499	0.283407	1.1912923	1.1912923
TCTAGCCTG	1.79E-05	1.73E-05	0.003271	0.561284	2.228265009	2.228265009
AACAGCCAC	5.33E-05	5.22E-05	0.002606	0.697098	2.621784062	2.621784062
CTAGCTGTC	2.08E-05	2.04E-05	0.001805	0.882905	3.154573595	3.154573595

Table 3. Frequencies of ninemers of in lean vs. obese samples with the highest significance (training and holdout sets: two halves of Study 1). The false discovery rate according to the Benjamini-Hochberg correction (shown in the last column) is rather high even if we only take the first fragment (about 10%).

Fragment	Male	Female	p (training set)	p (holdout set)	p (corrected)	FDR
TAGTACTGG	2.75E-05	2.85E-05	0.006019	0.174548	12.47301832	12.47301832
TTCATAGGG	3.39E-05	3.48E-05	0.0005157	0.305204	10.90478001	12.47301832
AGTCTCAGG	2.31E-05	2.23E-05	0.007333	0.353644	8.423677327	12.47301832
GATGTGTCT	3.88E-05	3.84E-05	0.006985	0.452399	8.081990361	12.47301832
GTCTCACAC	1.64E-05	1.59E-05	0.00236	0.49514	7.076437759	12.47301832
CTCAGTCT	0.0001047	0.0001014	0.006424	0.512597	6.104941306	12.47301832
CATGTAACC	2.97E-05	2.93E-05	0.001608	0.515833	5.26584129	12.47301832
GCTTCAGAC	4.10E-05	3.98E-05	0.006813	0.546829	4.884478864	12.47301832
CTCTAACAC	2.15E-05	2.10E-05	0.006313	0.578498	4.593207186	12.47301832
ACAGACTCA	3.89E-05	3.82E-05	0.007392	0.582096	4.159597401	12.47301832
GGTCAATTC	4.22E-05	4.27E-05	0.006413	0.59576	3.870217202	12.47301832
TGTGAGTCT	2.25E-05	2.20E-05	0.007573	0.618236	3.681541731	12.47301832
CAGACTCAT	4.51E-05	4.43E-05	0.007669	0.619291	3.404145383	12.47301832
GTGTTAGAC	1.63E-05	1.60E-05	0.004958	0.625175	3.191025321	12.47301832
ACCTCTGTC	4.03E-05	3.96E-05	0.005543	0.72925	3.474096374	12.47301832
GTCTAACAC	1.63E-05	1.60E-05	0.002582	0.752233	3.359611679	12.47301832
AGGATGTGT	4.81E-05	4.73E-05	0.001627	0.79598	3.345876584	12.47301832
TCTCCTCAA	5.78E-05	5.65E-05	0.006681	0.909561	3.61090455	12.47301832
TCTCAGTCT	3.36E-05	3.26E-05	0.004097	0.945748	3.556956171	12.47301832
GGTGTGTCT	2.86E-05	2.79E-05	0.005231	0.949554	3.392707002	12.47301832

Table 4. Frequencies of ninemers and an eightmer in female vs. male samples with the highest significance (training and holdout sets: two halves of Study 1). After the Benjamini-Hochberg correction, no significant differences can be found, as all FDR estimates are larger than 1.

6 Brain and graph theory

6.1 Background

Several large-scale projects for brain–mapping are being executed [89, 103], but the neuron-scale graph of the human brain, where the nodes are the neurons, and two neurons are connected by an edge if they are joined through a synapse, is out of reach today [166]. The difficulties come from the number of the neurons to be mapped, and also from the lack of the high-throughput methods for mapping their connections. The neuron-scale graphs were constructed only for very simple organisms with a very small number of neurons [38, 44, 145] or for just small cortical areas of more complex organisms [11, 60].

The application of magnetic resonance imaging (MRI) offers numerous methods for mapping physical and functional connections between subdivided anatomical areas of the brain (called "Regions of Interests", ROIs), each consisting of millions of neurons. The vertices are the ROIs, and two ROIs are connected by an edge if connections are detected between them by an MRI-based method. This method can either be diffusion MRI imaging, depicting the Brownian motion of water molecules in axons, consequently, mapping the axons between different cortical areas; or functional MRI (fMRI) imaging, depicting brain areas of elevated blood flow while the subject rests or performs different mental tasks.

In the present section we examine brain graphs (also called *connectomes*), computed from MRI scans originating from the data of the Human Connectome Project, recorded from male and female subjects between ages 22 and 35. Using advanced segmentation, parcellation and tractography algorithms one can produce graphs from good quality diffusion MRI imagery. We can then approach these graphs from two different perspectives: we can look for similarities between them and try to construct an "average" connectome, or we can explore the differences between the brain graphs and the reasons for them.

First we developed the Budapest Connectome Server, which is a web application

for generating, displaying and downloading an average brain graph from the individual graphs. We used the connectomes of 477 subjects to generate this average graph. This average graph is in fact configurable, which means that the user can set filter parameters for the edges like the edge weight combination mode, or the minimum number of subjects they need to appear in to be included in the consensus connectome.

Then we used demographic parameters (age and sex) to examine the differences between connectomes. After analyzing the graphs, no significant differences were found among age groups, which indicates that the brain graph of an individual may not change much over time, at least not during the age period 22–35. However, we found several significant differences between the male and female structural brain graphs. We show that the average female connectome has more edges, is a better expander graph, has larger minimal bisection width, and has more spanning trees than the average male connectome. Since the average female brain weighs less than the brain of males, these properties show that the female brain is more "well-connected" or perhaps, more "efficient" in a sense than the brain of males. We have found that the minimum bisection width, normalized with the edge number, is also significantly larger in the right and the left hemispheres in females: therefore, that structural difference is independent from the difference in the number of edges.

6.2 Data source and graph construction

The dataset applied is a subset of the Human Connectome Project [103] anonymized 500 Subjects Release:

(http://www.humanconnectome.org/documentation/S500) of healthy subjects between 22 and 35 years of age. Data was downloaded in October, 2014. The Connectome Mapper Toolkit [42] (http://cmtk.org) was applied for brain tissue segmentation into grey and white matter, partitioning, tractography and the construction of the graphs from the fibers identified in the tractography step. The Connectome Mapper Toolkit [42] default partitioning was used, computed by the FreeSurfer, and

based on the Desikan-Killiany anatomical atlas. Using all the 5 different resolution options, we partitioned the gray matter into 83, 129, 234, 463 and 1015 cortical and sub-cortical structures (as the brainstem and deep-grey nuclei), referred to as "Regions of Interest", ROIs, (see Figure 4 in [42]). Tractography was performed using the deterministic streamline method [42] with randomized seeding.

The graphs were constructed as follows: the nodes correspond to the ROIs in the specific resolution. Two nodes were connected by an edge if there exists at least one fiber (determined by the tractography step) connecting the ROIs, corresponding to the nodes. More than one fibers, connecting the same nodes, may give rise to the weight of that edge, depending on the weighting method. Loops were deleted from the graph.

The weights of the edges are assigned by several methods, taking into account the lengths and the multiplicities of the fibers, connecting the nodes:

- `Unweighted`: Each edge has weight 1.
- `FiberN`: The number of fibers traced along the edge: this number is larger than one if more than one fibers connect two cortical or sub-cortical areas, corresponding to the two endpoints of the edge.
- `FAMean`: The arithmetic mean of the fractional anisotropies [20] of the fibers, belonging to the edge.
- `FiberLengthMean`: The average length of the fibers, connecting the two endpoints of the edge.
- `FiberNDivLength`: The number of fibers belonging to the edge, divided by their average length. This quantity is related to the simple electrical model of the nerve fibers: by modeling the fibers as electrical resistors with resistances proportional to the average fiber length, this quantity is precisely the conductance between the two regions of interest. Additionally, `FiberNDivLength` can be observed as a reliability measure of the edge: longer fibers are less reliable than the shorter ones, due to possible error accumulation in the tractography

algorithm that constructs the fibers from the anisotropy data. Multiple fibers connecting the same two ROIs, corresponding to the endpoints, add to the reliability of the edge, because of the independently tractographed connections.

6.3 The Budapest Connectome Server

The connectomes of different human brains are pairwise distinct: we cannot talk about an abstract "graph of the brain". Two typical connectomes, however, have quite a few common graph edges that may describe the same connections between the same cortical areas. The Budapest Reference Connectome Server v3.0 generates the common edges of the connectomes of 477 distinct cortexes, each with 1015 vertices, computed from 477 MRI data sets of the Human Connectome Project. The user may set numerous parameters for the identification and filtering of common edges, and the graphs are downloadable in both csv and GraphML formats; both formats carry the anatomical annotations of the vertices, generated by the Freesurfer program. The resulting consensus graph is also automatically visualized in a 3D rotating brain model on the website.

The consensus graphs, generated with various parameter settings, can be used as reference connectomes based on different, independent MRI images, therefore they may serve as reduced-error, low-noise, robust graph representations of the human brain. The Budapest Connectome Server is available for users at the URL http://connectome.pitgroup.org .

In this section we present this webserver, which, starting from the diffusion MRI data published as a result of the Human Connectome Project [103], compiles differently parametrized, customizable reference graphs from the common edges of the graphs describing 477 different 1015-vertex graphs of 477 human subjects. For users who would like to use the default settings without customization, a single graph, the Budapest Reference Connectome v3.0 is also presented in two downloadable formats on the server page. The default, canonical "Budapest Reference Connectome v3.0" can be downloaded by simply hitting the "Download graph" button without

changing the default options. This default graph has 1015 vertices and 1000 edges.

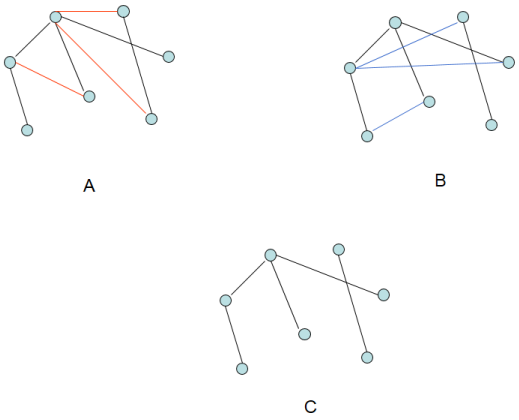

Figure 11: Some edges of graphs A and B are common; they form graph C, the consensus graph.

The resulting graphs may be used for identifying more robust, more error-free connections between the cortical areas, represented by ROIs: for example, in the default reference graph (i.e., the Budapest Reference Connectome v3.0), if an edge is present then it is present in at least 209 different source graphs (50% of them). In general, one may set the "Minimum edge confidence" to value k anywhere between $k = 1$ (where an edge is included if it is present in at least one source graph) through $k = 477$ (where an edge is present in the reference graph if it can be found in all the 477 source graphs).

Therefore, the resulting graphs contain *common, consensus* edges (i.e., Fig. 11) originating from multiple graphs, computed from the diffusion MRI data of different subjects.

Version 3.0 of the Budapest Reference Connectome Server is described here in detail. Choosing Version 1.0/2.0 is also possible on the website: Version 1.0 applies the source data from the classical article of [66] describing six connectomes of five subjects, each with 998 vertices; while Version 2.0 includes 96 graphs (a subset of Version 3.0).

By filtering edges with very few occurrences or those with small weights, one

may get a connectome with more reliable edges and weights than in the case of any single dataset in the input. Therefore, we may get a robust large-scale weighted graph model of the human brain through the consensus graphs generated by the server.

6.3.1 Compilation

The source dataset was a subset of Human Connectome Project 500 Subjects Release (http://www.humanconnectome.org/documentation/S500/), containing MRI images of healthy adult males and females between the ages of 22 and 35. Data were downloaded in October, 2014. Partitioning, tractography, and graph construction were done by the Connectome Mapper Toolkit (http://cmtk.org). Partitioning of the gray matter was done by the Lausanne2008 method [66] into 1015 ROIs. For tractography, the so-called deterministic streamline method was applied.

The graphs were constructed as follows: the nodes correspond to the 1015 ROIs. Two nodes are connected by an edge if there exists at least one fiber that connects these two ROIs. The number or length of the fibers connecting these nodes will define the weight of that edge, depending on the weight function. We deleted all loops from the graph.

After 477 graphs were computed, each with 1015 vertices, we identified the common edges, their confidence (the number of graphs that include that edge), and weights, computed according to their median and mean. The large, pre-computed tables were integrated into the webserver.

Version 1.0 of the webserver applies the six graphs that were described in [66]. The definition of *weight* (also called *strength*) and its computation, and also the parcellation of the cortex used are described in [66]. The six connectomes were downloaded from http://www.cmtk.org/datasets/homo_sapiens_01.cff in September, 2014.

The visualization component employs a heavily modified version of the WebGL Brain Viewer [61]. We transformed the viewer module so that in can be used in window mode instead of full-screen mode. We made the nodes of the graph clickable.

After clicking a ROI, its neighbors and connections are displayed. The user has to click on the background to see the whole graph again. Hovering over a node displays the name of the corresponding ROI in a tooltip. We also changed the drawing order of the graph and the brain surface overlay, so now the edges are more clearly visible. We also regenerated the 3D node positions – this was needed because the subcortical nuclei originally had the same coordinate so they could not be distinguished when displaying the graph.

6.3.2 User interface and operation

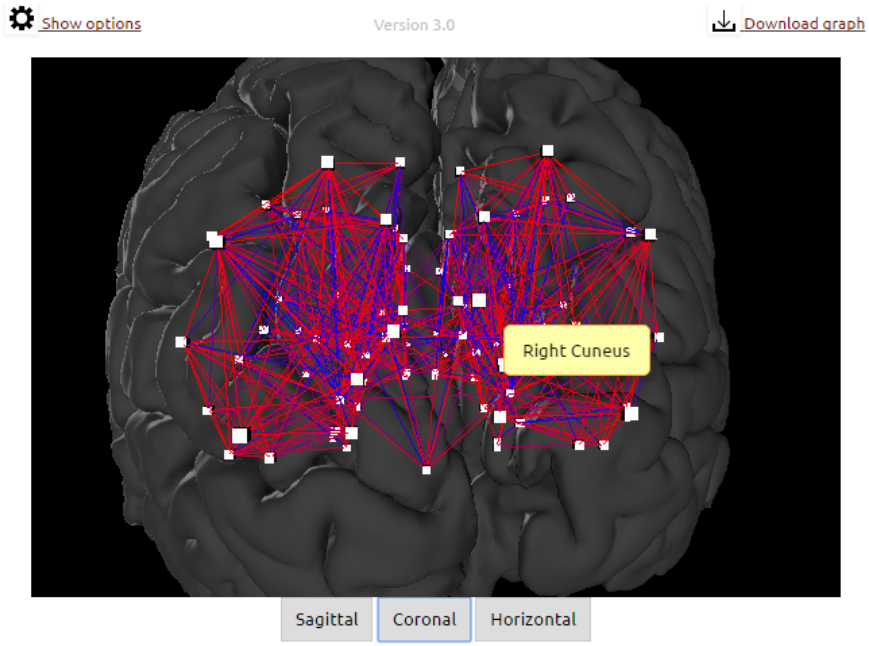

Figure 12: The Budapest Reference Connectome server (coronal view)

After opening the Budapest Connectome Server website and choosing the "Show options" button, the following options can be set:

(i) Version 1.0, 2.0 or 3.0. The default choice is 3.0, using the graphs of 477 subjects, computed from the Human Connectome Project [103]. The user may alternatively choose Version 1.0 that has only six graphs computed and described by the classical article of [66], or Version 2.0 having 96 graphs computed by us from a subset of the MRI images.

(ii) Population: It is possible to see the female and male consensus brain graph, or the consensus graph for the whole population (default option).

(iii) Minimum edge confidence: The graph to be constructed will contain all the edges that are present in at least k graphs, connecting the very same vertices in each graph. Valid choices are $k = 1 \dots 477$. For example, $k = 477$ means that an edge is presented in the resulting consensus graph if and only if each source graph contains that edge.

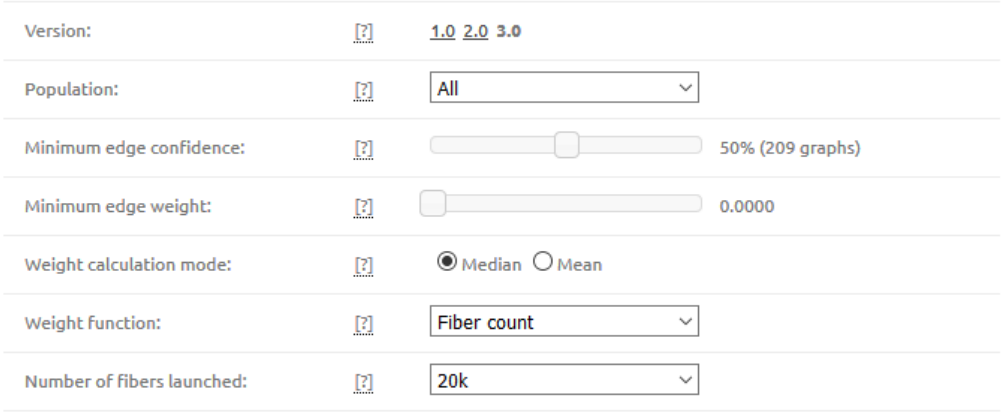

Figure 13: The Budapest Reference Connectome server, options panel

For each edge $\{u, v\}$, the weight of that edge can be defined in multiple different ways. Budapest Connectome Server offers 4 options: electrical connectivity, fiber count, average fiber length and fractional anisotropy. The *electrical connectivity* is defined here as a fraction n/L, where n is the number of fibers connecting u and v, and L is the average length of the fibers. It may or may not correspond to actual electrical connectivity.

(iv) Minimum edge weight: One may set a slider to a value of minimum weight required. The returned graph will contain edges whose mean or median weights are larger than or equal to this value. The mode of computation (mean or median) can be set by another option.

(v) Weight calculation mode: There are two choices: Median or Mean. The default choice is the median, since the median is more robust and error-prone than

Label	Description
id_node1	the numerical ID of the first vertex of the edge
id_node2	the numerical ID of the second vertex of the edge
name_node1	the anatomical name of node 1
name_node2	the anatomical name of node 2
parent_id_node1	the ID of the parent region of node 1 on the 83-region atlas
parent_id_node2	the ID of the parent region of node 2 on the 83-region atlas
parent_name_node1	the name of the parent region of node 1 on the 83-region atlas
parent_name_node2	the name of the parent region of node 2 on the 83-region atlas
minimum_edge_confidence	the number of the graphs in which the edge is contained
median	the median of the weights of the same edge in different graphs
average	the average of the weights of the same edge in different graphs

Table 5: The column labels of the result file in csv format. The 83-region atlas refers to the atlas of the FreeSurfer tool.

the mean: extremely large or small edge strengths typically have less impact to the median than to the mean.

(vi) Weight function: Edge weight calculation mode: electrical connectivity, fiber count, average fiber length or fractional anisotropy.

(vii) Number of fibers launched: we ran the tractography with different options (20,000, 200,000 or 1,000,000 fibers launched), the number of fibers traced can be selected here.

The resulting graph can be downloaded in CSV or GraphML formats, or can readily be visualized on the webpage. For better reproducibility, the downloaded filenames contain the parameter settings as follows: the default graph (Budapest Reference Connectome Version 3.0) will be downloaded as a comma-separated file named `budapest_connectome_3.0_209_0_median.csv`. That is, the csv file contains the graph generated by Version 3.0 of the server, with a minimum confidence of 209 (i.e., each edge of the graph is contained in at least 209 input graphs), a minimum edge weight of 0, and the weights of the edges of the reference graph are computed as the median of the weights of the corresponding edges of the input graphs.

The columns of the CSV file are described in the table above.

We also compiled a chart about the number of common edges in at least n graphs (n=1,2,...,477). This chart is shown in Figure 14.

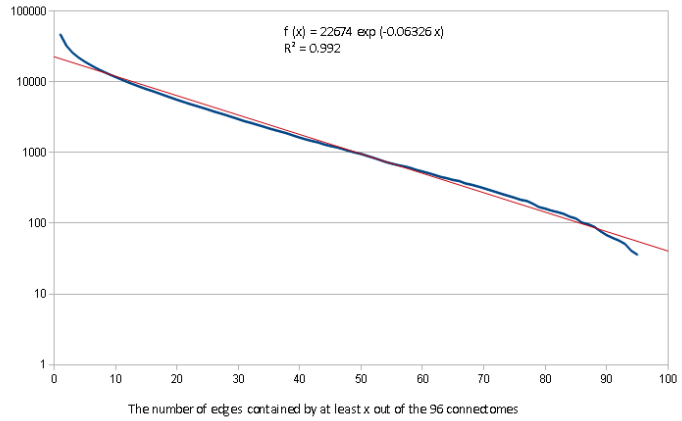

Figure 14: Plot of the number of common edges.

6.4 Comparative connectomics

In the last several years hundreds of publications were published describing or analyzing structural or functional networks of the brain, frequently referred to as "connectome" [41, 67]. Some of these publications analyzed data from healthy humans [18, 19, 21, 64], and some compared the connectome of the healthy brain with the diseased one [5, 7, 17, 26, 28].

So far, connectome analysis mostly used tools developed for very large networks of very similar agents, such as the graph of the World Wide Web (with billions of vertices), or protein-protein interaction networks (with tens or hundreds of thousands of vertices), and because of the huge size of original networks, these methods used only very fast algorithms and frequently just primary degree statistics and graph-edge counting between pre-defined regions or lobes of the brain [83]. But this is not the only problem with this approach. We think that the connectome should not be viewed as a self-organizing network like Barabási graphs, because the brain consists of areas with a highly specialized function, while "classical" networks consist of lots of very similar agents. While it certainly makes sense to investigate the graph constituted by these regions with network-theoretical means, it is more advisable to

view it as a block diagram of a machine with functional modules, and focus on how the graph helps the transmission of information between these functional regions.

Deep graph-theoretic ideas in the context with the graph of the World Wide Web led to the definition of Google's PageRank and the subsequent rise of the most-popular search engine to date. Brain graphs, or connectomes, are being widely explored today. We believe that non-trivial graph theoretic concepts, similarly as it happened in the case of the World Wide Web, will lead to discoveries enlightening the structural and also the functional details of the animal and human brains. When scientists examine large networks of tens or hundreds of millions of vertices, only fast algorithms can be applied because of the size constraints. In the case of diffusion MRI-based structural human brain imaging, the effective vertex number of the connectomes, or brain graphs derived from the data is on the scale of several hundred today. That size facilitates applying strict mathematical graph algorithms even for some hard-to-compute (or NP-hard) quantities like vertex cover or balanced minimum cut.

We demonstrated that deep and more intricate graph theoretic parameters could also be computed by using, among other tools, contemporary integer programming approaches for connectomes with several hundred vertices, and they can be used to explore differences between female and male connectomes. With these mathematical tools we showed statistically significant differences in some graph properties of the connectomes, computed from MRI imaging data of male and female brains. We will not try to associate behavioral patterns of males and females with the discovered structural differences [83] (see also the debate that article has generated: [55,84,88]), because we do not have behavioral data of the subjects of the imaging study, and, additionally, we believe that one cannot describe high-level functional properties implied by those structural differences. However, we clearly demonstrate that, in an information-theoretical sense, deep graph-theoretic parameters show "better" connections in female connectomes than in male ones.

The study of [83] analyzed the 95-vertex graphs of 949 subjects aged between 8 and 22 years, using basic statistics for the numbers of edges running either between or

within different lobes of the brain (the parameters deduced were called *hemispheric connectivity ratio, modularity, transitivity and participation coefficients*, see [83] for the definitions). It was found that males have significantly more intra-hemispheric edges than females, while females have significantly more inter-hemispheric edges than males.

We analyzed the connectomes of 96 subjects, 52 females and 44 males, each with 83, 129 and 234 node resolutions and five distinct arc weight functions. We considered the connectomes as graphs with weighted edges, and performed graph-theoretic analyses with computing some polynomial-time computable and also some NP-hard graph parameters on the individual graphs, and then compared the results statistically for the male and the female group.

We have found that female connectomes have more edges, larger (normalized) minimum bisection widths, larger minimum-vertex covers and more spanning trees than the male connectomes.

6.4.1 Graph parameters

By *generalized adjacency matrix* we mean a matrix of size $n \times n$ where n is the number of nodes (vertices) in the graph, whose rows and columns correspond to the nodes, and whose elements are either zero if there is no edge between the two nodes, or equals to the weight of the edge connecting the two nodes. By the *generalized degree* of a node we mean the sum of the weights of the edges adjacent to that node. Note that the generalized degree of the node v is exactly the sum of the elements in the row (or column) of the generalized adjacency matrix corresponding to v. By *generalized Laplacian matrix* we mean the matrix $D - A$, where D is a diagonal matrix containing the generalized degrees, and A is the generalized adjacency matrix.

We calculated various graph parameters for each brain graph and weight function. These parameters included:

- Number of edges (`Sum`). The weighted version of this quantity is the sum of the weights of the edges.

- Normalized largest eigenvalue (`AdjLMaxDivD`): The largest eigenvalue of the generalized adjacency matrix, divided by the average generalized degree. Dividing by the average degree of vertices was necessary because the largest eigenvalue is bounded by the average and maximum degrees, and thus is considered by some a kind of "average degree" itself [97]. This means that a denser graph may have a bigger λ_{max} largest eigenvalue solely because of a larger average degree. We note that the average degree is already defined by the sum of weights.

- Eigengap of the transition matrix (`PGEigengap`): The transition matrix P_G is obtained by dividing all the rows of the generalized adjacency matrix by the generalized degree of the corresponding node. When performing a random walk on the graph, for nodes i and j, the corresponding matrix element describes the probability of transitioning to node j, supposing that we are at node i. The eigengap of a matrix is the difference of the largest and the second largest eigenvalue. It is characteristic to the expander properties of d-regular graphs: the larger the gap, the better expander is the graph (see [76] for the exact statements and proofs).

- Hoffman's bound (`HoffmanBound`): The expression

$$1 + \frac{\lambda_{max}}{|\lambda_{min}|},$$

  where λ_{max} and λ_{min} denote the largest and smallest eigenvalues of the adjacency matrix. It is a lower bound for the chromatic number of the graph. The chromatic number is generally higher for denser graphs, as the addition of an edge may make a previously valid coloring invalid.

- Logarithm of number of spanning forests (`LogAbsSpanningForestN`): The number of the spanning trees in a connected graph can be calculated from the spectrum of its Laplacian [39, 94]. Denser graphs tend to have more spanning trees, as the addition of an edge introduces zero or more new spanning trees.

If a graph is not connected, then the number of spanning forests is the product of the numbers of the spanning trees of the components. The parameter `LogAbsSpanningForestN` equals to the logarithm of the number of spanning forests in the unweighted case. In the case of other weight functions, if we define the weight of a tree by the product of the weights of its edges, and the weight of a forest by the product of the weights of its components, then this parameter equals to the logarithm of the sum of the weights of the forests.

- Balanced minimum cut, divided by the number of edges (`MinCutBalDivSum`): The task is to partition the graph into two sets whose size may differ from each other by at most 1, so that the number of edges crossing the cut is minimal. This is the "balanced minimum cut" problem, or sometimes called the "minimum bisection width" problem. For the whole brain graph, our expectation was that the minimum cut corresponds to the boundary of the two hemispheres, which was indeed proven when we analyzed the results.

- Minimum cost spanning forest (`MinSpanningForest`), calculated with Kruskal's algorithm.

- Minimum weighted vertex cover (`MinVertexCover`): Each vertex should have a (possibly fractional) weight assigned such that, for each edge, the sum of the weights of its two endpoints is at least 1. This is the fractional relaxation of the NP-hard vertex-cover problem [73]. The minimum of the sum of all vertex-weights is computable by a linear programming approach.

- Minimum vertex cover (`MinVertexCoverBinary`): Same as above, but each weight must be 0 or 1. In other words, a minimum size set of vertices is selected such that each edge is covered by at least one of the selected vertices. This NP-hard graph-parameter is computed only for the unweighted case. The exact values are computed by an integer programming solver SCIP (http://scip.zib.de), [1, 2].

The above 9 parameters were computed for all three resolutions and for the left

and the right hemispheres and also for the whole connectome, with all 5 weight functions (with the following exceptions: `MinVertexCoverBinary` was computed only for the unweighted case, and the `MinSpanningForest` was not computed for the unweighted case).

6.4.2 Statistical analysis

Since each connectome was computed in multiple resolutions (in 83, 129 and 234 nodes), we had three graphs for each brain. There were in fact 5 versions of each graph, but we did not use the two highest resolutions, only the 3 anatomically more robust ones, with fewer nodes. In addition, the parameters were calculated separately for the connectome within the left and right hemispheres as well, not only the whole graph, since we intended to examine whether statistically significant differences can be attributed to the left or right hemispheres. Each subjects' brain was corresponded to 9 graphs (3 resolutions, each in the left and the right hemispheres, plus the whole cortex with sub-cortical areas) and for each graph we calculated 9 parameters, each (with the exceptions noted above) with 5 different edge weights. This means that we assigned $7 \cdot 5 \cdot 3 + 1 \cdot 3 + 4 \cdot 3 = 120$ attributes to each resolution of the 96 brains, that is, 360 attributes to each brain.

The statistical null hypothesis [74] of ours was that the graph parameters do not differ between the male and the female groups. As the first approach, we have used ANOVA (Analysis of variance) [157] to assign p-values for all parameters in each hemispheres and in each resolutions and in each weight-assignments.

Our very large number of attributes may lead to false negatives, i.e., to "type II" statistical errors: in other words, it may happen that an attribute, with a very small p-value may appear at random, simply because we tested a lot of attributes. In order to deal with type II statistical errors, we followed the route described below.

We divided the population randomly into two sets by the parity of the sum of the digits in their ID. The first set was used for making hypotheses and the second set for testing these hypotheses. This was necessary to avoid type II errors resulting from

multiple testing correction. If we made hypotheses for all the numerical parameters, then the Holm-Bonferroni correction [75] we used would have unnecessarily increased the p-values. Thus we needed to filter the hypotheses first, and that is why we needed the first set. Testing on the first set allowed us to reduce the number of hypotheses and test only a few of them on the second set.

The hypotheses were filtered by performing ANOVA (Analysis of variance) [157] on the first set. Only those hypotheses were selected to qualify for the second round where the p-value was less than 1%. The selected hypotheses were then tested for the second set as well, and the resulting p-value corrected with the Holm-Bonferroni correction method [75] with a significance level of 5%.

In Table 1 those hypotheses rejected were highlighted in bold, meaning that *all* the corresponding graph parameters differ significantly in sex groups at a combined significance level of 5%.

We also highlighted (in italic) those p-values which were individually less than the threshold, meaning that these hypotheses can *individually* be rejected at a level of 5%, but it is very likely that some of these graph parameters are in fact not significantly different between the sexes.

6.4.3 Results and discussion

For a list of all the significantly different graph parameters, see 8.2 in the Appendix.

In order to describe the parameters that differ significantly among male and female connectomes, we need to place them in the context of their graph theoretical definitions.

6.4.3.1 Edge number and edge weights

We have found a significantly higher number of edges in females. We counted the edges with 5 different weight functions and also without any weights) in both hemispheres and also in the whole brain, in all resolutions, and almost all configurations

resulted in a significant difference between the two sexes, women having more edges in their brain graph. This finding is surprising, since we used the same parcellation, the same tractography and the same graph construction methods for female and male brains, and because it is proven that the brains of females weigh less than those of males on average [156]. For example, in the 234-vertex resolution, the average number of (unweighted) edges in female connectomes is 1826, in males 1742, with $p = 0.00063$ (see the Appendix for tables with the results). The work of [83] reported similar findings in inter-hemispheric connections only.

6.4.3.2 Minimum cut and balanced minimum cut

Suppose that the nodes (vertices) of a graph are partitioned into two disjoint non-empty sets, say X and Y; their union is the whole vertex set of the graph. The X, Y *cut* is the set of *all* the edges connecting a vertex in X with a vertex in Y (Figure 15A). The *size* of the cut is the number of edges in the cut, or the sum of the weights of the edges if the graph has an associated edge weight function. The *minimum cut* between vertices a and b is the minimum cut, taken for all X and Y, where vertex a is in X and b is in Y. This quantity gives the "bottleneck", in a sense, between those two nodes (c.f., Menger theorems and the Ford-Fulkerson MFMC theorem [57, 96]). The *minimum cut in a graph* is defined to be the cut with the fewest edges for *all* non-empty sets X and Y, partitioning the vertices.

Clearly, for non-negative weights, the size of the minimum cut in a disconnected graph is 0. On the other hand, in connected graphs, the minimum cut is very frequently determined by just the node with the smallest degree: that node is the only element of set X and all the other vertices of the graph are in Y (Figure 15B). Because of this observed phenomenon, the minimum cut is frequently queried for the "balanced" case, when the size (i.e., the number of vertices) of X and Y needs to be equal (or, more exactly, may differ by at most one if the number of the vertices of the graph is odd), see Figure 15C. This problem is referred to as *the balanced minimum cut* or the *minimum bisection* problem. We say that the minimum bisection is *small* in a sense if there exists a partition of the vertices into two sets of equal size that

are connected with only a few edges. If the minimum bisection is large then the two half-sets in *every possible bisections* of the graph are connected by many edges.

Therefore, the balanced minimum cut of a graph is independent of the particular labeling of the nodes. The number of all the balanced cuts in a graph with n vertices is greater than

$$\frac{1}{n+1}2^n,$$

that is, for $n = 250$, this number is very close to the number of atoms in the visible universe [3], so a brute force approach certainly does not work for the minimum bisection problem. The complexity of computing this quantity is known to be NP-hard [59] in general, but with contemporary integral programming approaches, for the graph-sizes we are dealing with, the exact values are computable.

In computer engineering, an important measure of the quality of an interconnection network is its minimum bisection width [144]: the bigger the minimum bisection width the better the network.

For the whole brain graph, as it had been anticipated, we found that the minimum balanced cut almost exactly represents the edges crossing the *corpus callosum*, connecting the two cerebral hemispheres. This means that the intrahemispheric connections are much more prevalent than the interhemispheric ones. In other words, the hemispheres naturally partition the brain into two rather dense subgraphs which are connected by only a relatively few number of edges.

We showed that within both hemispheres, the minimum bisection size of female connectomes are significantly larger than the minimum bisection size of the males. Much more importantly, we show that this remains true if we *normalize with the sum of all edge-weights*: that is, this phenomenon cannot be due to the higher number of edges or the greater edge weights in the female brain: it is an intrinsic property of the female brain graph in our data analyzed.

For example, in the 234-vertex resolution, in the left hemisphere, the normalized balanced minimum cut in females, on the average, is 0.09416, in the males 0.07896, $p = 0.00153$ (see the Appendix for tables with the results).

We think that this finding is one of the main results of the present work: even if the significant difference in the weighted edge numbers are due to some artifacts in the data acquisition/processing workflow, the normalized balanced minimum cut size seems to be independent from those processes.

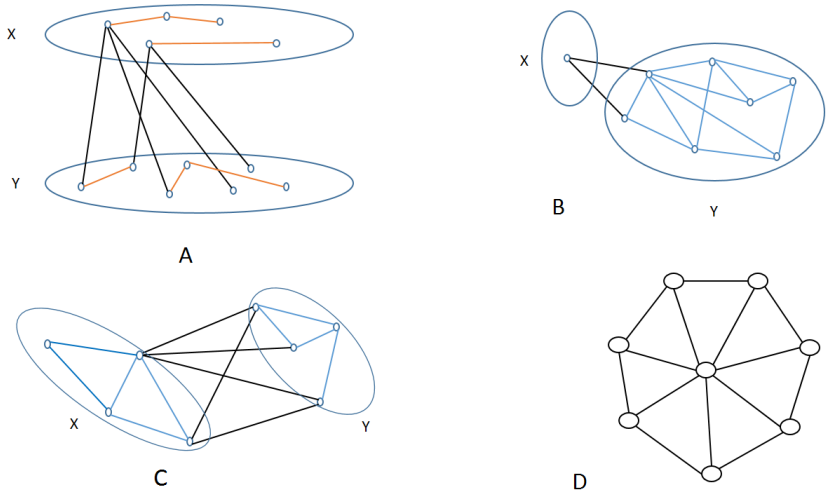

Figure 15: Panel A: An X-Y cut. The cut edges are colored black. Panel B: An unbalanced minimum cut. Panel C: A balanced cut. Panel D: The wheel graph.

6.4.3.3 Eigengap and the expander property

Expander graphs and the expander property of graphs are one of the most interesting area of graph theory: they are closely related to the convergence rate and the ergodicity of Markov chains, and have applications in the design of communication- and sorting networks and methods for de-randomizing algorithms [76]. A graph is an ε-expander, if every — not too small and not too large — vertex-set S of the graph has at least $\varepsilon|S|$ outgoing edges [76]. In mathematical terms,

$$|\partial S| \geq \varepsilon|S|, \forall 0 < |S| \leq n/2,$$

where n is the number of vertices and ∂S is set of edges crossing the vertex S.

Random walks on good expander graphs converge very fast to the limit distribution: in lay terms this means that good expander graphs, in a certain sense, are "intrinsically better" connected than bad expanders. It is known that a large eigengap of the walk transition matrix of a graph implies good expansion property [76].

We have found that women's connectomes have a significantly larger eigengap, and, consequently, they are better expander graphs than the connectomes of men. For example, in the 83-node resolution, in the left hemisphere and in the unweighted graph, the average female connectome's eigengap is 0.306 while in the case of men it is only 0.272, with $p = 0.00458$. It is possible that this is a consequence of a higher overall number of edges.

6.4.3.4 The number of spanning forests

A graph-theoretical *tree* is a connected cycle-free graph. All trees on n vertices has $n - 1$ edges. Trees and tree-based structures are common in natural sciences: phylogenetic trees, hierarchical clusters, data-storage on hard-disks, or a computational model called *decision trees* all apply graph-theoretical trees. A *spanning forest* is a minimal subgraph of a graph which has the same number of components as the containing graph, and a *spanning tree* is a minimal subgraph of a connected graph that is still connected. Clearly, a tree has only one spanning tree (itself). Any connected graph on n vertices has a minimum of $n - 1$ and a maximum of $n(n - 1)/2$ edges [97]. A connected graph with few edges still may have exponentially many distinct spanning trees: e.g., the n-vertex wheel on Figure 15D has at least 2^{n-1} spanning trees (for $n \geq 4$). Cayley's famous theorem, and its celebrated proof with Prüfer codes [114] shows that the number of spanning trees of the complete graph on n vertices is n^{n-2}.

If a graph is not connected, then it contains more than one connected components. Each connected component has at least one spanning tree, and the whole graph has at least one spanning forest, the union of the spanning trees of the components. The number of spanning forests is clearly the product of the numbers of

the spanning trees of the components. For graphs in general, one can compute the number of their spanning forests by Kirchoff's matrix tree theorem [39,94] using the eigenvalues of the Laplacian matrix [39] of the graph.

We show that female connectomes have a significantly higher number of spanning trees than the connectomes of males. For example, in the 129-vertex resolution, in the left hemisphere, the logarithm of the number of the spanning forests in the unweighted case are 162.01 in females, 158.88 in males with $p = 0.013$. This is not surprising, since the brain graphs of females have more edges in general.

6.4.4 Conclusion

To sum up, we have computed 83-, 129- and 234-vertex-graphs from the diffusion MRI images of the 96 subjects of 52 females and 44 males, between the age of 22 and 35. We have found, after a careful statistical analysis, significant differences between some graph theoretical parameters of the male and female brain graphs. Our findings show that the female brain graphs have generally more edges (counted with and without weights), have larger normalized minimum bisection widths and have more spanning forests (counted with and without weights) than the connectomes of males (Table 1). Additionally, with weaker statistical validity, some spectral properties and the minimum vertex cover also differ in the connectomes of different sexes (each with $p < 0.02$).

7 Correlations, maximum spanning trees and the Human Connectome Project

7.1 Introduction

A large amount of human psychological and behavioral data were collected and deposited in the last several decades worldwide. In the framework of the Human Connectome Project [103] those type of data were enriched with structural and functional MR images of the same subjects. In the present section, we are analyzing the graph-theoretical properties of the connectomes as well as the psychological- and behavioral test data that were published in the Human Connectome Project's [103] anonymized 500 Subjects Release. Our goal is finding correlations between those highly inhomogeneous data items. We are considering Pearson's product-moment correlation, which well describes the linear connections between attributes, and also Spearman's rank correlation, which well describes non-linear connections between attributes [127].

Some of the strongest correlations are natural, describing closely related quantities (e.g., the volume and the relative volume of the same brain area, or graph maximum matching numbers and minimum vertex covers). Some of them are novel, and are detailed in this section, and some of them were discovered in the recent years (e.g., the connection between gambling behavior and the number of connections crossing the insular cortex [49]).

7.2 Maximum weight spanning trees of correlations

Suppose we have a large number of attributes, describing the properties of a complex system. Frequently, a straightforward step in their analysis is the computation of the pairwise correlations of the attributes. If we have n attributes, then one can form $n(n-1)/2$ pairs from them, so we will need to compute that many pairwise

correlations as well. Identifying the most "important" correlations from the set is, generally, not an easy task.

A possible natural requirement is generating a "non-redundant set" of correlations in the following sense: Suppose that random variable A strongly correlates with B, and random variable B strongly correlates with C, then, usually, A and C are also strongly correlated. Now, if we want to find a non-redundant set of correlations between A, B and C, it is an obvious idea to choose the two strongest correlation between them, say between A and B and between B and C, and to leave out the weakest, say the one between C and A. We can visualize those non-redundant "strong"' or "important" correlations by a graph on vertices A, B and C, with two edges: AB and AC.

In the general case when n attributes are considered, we are interested in cycle-free, connected graphs with the highest possible total weight on its edges, corresponding to the absolute values of the correlations. The cycle-free property ensures the non-redundancy, and the highest total weight of the absolute values of the correlations ensures that we have chosen the "most relevant" ones.

The idea of constructing the maximum weight spanning tree from the pairwise correlation coefficients has been applied before in several settings.

Mantegna [101] constructed a graph from financial equities, traded in stock markets, and weighted the edges of the graph by the correlations computed from the time series of the logarithms of the stock prices. It was found that – essentially – a maximum-weight spanning tree well-describes several known relations and suggests numerous new ones between the time series. In [30] Section 9.3.5 and [71] the maximum-weight spanning trees are computed explicitly in similar settings.

In [168] correlations related to the co-expression of gene-pairs of the yeast (*S. cerevisiae*) were computed, and a graph was constructed with the genes as the vertices and the co-expression correlation-weighted edges for each pair of genes. Then a maximum weight spanning tree was computed and visualized for demonstrating a non-redundant system of strong correlations between the genes. ([168], Fig. 6).

More recently, in [65], the maximum-weight spanning tree of the correlations was applied for feature selection in weakly-structured multimedia data. In [32] a similar method is used for finding related attributes in a regional Italian hospitality industry.

7.3 Materials and Methods

Our data source was the Human Connectome Project's [103] anonymized 500 Subjects Release. In the dataset diffusion and functional MRI data, psychological test results, and some cognitive data have been published. Here we list the different types of data applied by us.

The subject data table from the Human Connectome Project consists of 527 rows (corresponding to subjects) and 451 attributes.

Diffusion MRI images of the subjects were processed by the researchers of the Human Connectome Project with the Freesurfer software suite [56] to obtain the size of various regions of interests (ROIs), i.e., the area, thickness and volume of major cortical and sub-cortical areas of the brain. For cortical regions, only average thickness and surface area were available, so we multiplied these two quantities to obtain the approximate volume of the ROI. We divided the volume of an ROI by `FS_Mask_Vol` (Freesurfer Brain Mask Volume, i.e., roughly the brain volume) to obtain the *relative volume* of that region. We intended to compensate for brain size because it is already well known that males on the average have larger brains than females [156]. We added these new, relativized attributes to the data table.

Several psychological and cognitive tests were also performed on the subjects. These included the MMSE (Mini Mental State Examination), various NIH Toolbox [150] cognitive tests (Picture Sequence Memory Test, Dimensional Change Card Sort Test, Flanker Inhibitory Control and Attention Test, Oral Reading Recognition Test, Picture Vocabulary Test, Pattern Comparison Processing Speed Test, List Sorting Working Memory Test), NIH Toolbox Emotion Domain (Anger-Affect, Anger-Hostility, Anger-Physical Aggression, Fear-Affect, Fear-Somatic Arousal, Sadness,

General Life Satisfaction, Meaning and Purpose, Positive Affect, Friendship, Loneliness, Perceived Hostility, Perceived Rejection, Emotional Support, Instrumental Support, Perceived Stress, Self-Efficacy), a test for self-regulation/impulsivity (Delay Discounting), Penn Line Orientation Test, Penn Continuous Performance Test, Penn Word Memory Test and Penn Emotion Recognition Test.

The subjects were also asked to perform some fMRI tasks, including identifying random and non-random shape movements, a working memory test (places, faces, body parts, and tools), language, math and gambling tasks.

We also added numerous attributes to the data table corresponding to various graph parameters of the connectomes of the subjects. These included the total number of the connectome edges, counted with weights, maximum matching and minimum vertex cover, Hoffman's bound (a bound for the chromatic number), the eigengap of the transition matrix (which is a quantity connected with the properties of random walks on the graph), and the total number of edges exiting different lobes of the brain. These graph-theoretical quantities of the connectomes were defined in details and also computed in the articles [139, 141].

We calculated the correlations between all possible pairs of the attributes (columns of the database). The goal was to obtain a non-redundant list of important correlations. Not surprisingly, our observation was that correlation is transitive in most of the cases, so if A correlates with B and B correlates with C, then this usually implies that A correlates with C in some degree. Therefore, if we consider a complete graph whose vertices are the attributes, and whose edges represent correlations between two attributes, then our goal can be reformulated as follows: we want to find a subgraph without cycles (because cycles usually mean redundant correlations), and whose edges correspond to relatively large correlations (because larger correlations are more important than the others).

This optimization problem is essentially a maximum weight spanning tree problem, which can be solved by well-known graph theoretical algorithms such as the Kruskal algorithm [96]. The classical question is finding a *minimum* weight spanning tree, but by a straightforward transformation, the algorithm can be used for

finding the maximum weight spanning tree. We applied this method to the HCP (Human Connectome Project) subject data table in its anonymized 500 Subjects Release [103] in order to search for connections between psychological and cognitive scores, demographic data, and brain ROI sizes.

The possible age groups of the subjects were 22-25, 26-30, 31-35 and 36+. Only 3 of the subjects were 36 years or older. We translated the age groups to numbers the following way: 0 meant 22-25, 1 meant 26-30, 2 meant 31-35 and 3 meant 36+. We translated the "gender" attribute to 1 (male) and 2 (female). We had to convert these attributes to numbers so we can calculate correlations between them and other attributes.

We have computed both the Pearson's product-moment correlation (this is "the correlation", most frequently computed in science), which well describes the linear connections between attributes, and also Spearman's rank correlation, which well describes non-linear connections between attributes [127].

7.4 Results

7.4.1 Maximum spanning tree of Pearson's correlations

The maximum spanning tree is visualized on Figure 16 and in a more viewable form in an interactive figure at `http://pitgroup.org/static/graphmlviewer/index.html?src=correl_spanning_tree.graphml`.

The spanning tree had 716 edges with non-zero correlation, and 717 attributes, so the graph contained 717 vertices. The weakest edge still had 15% correlation.

The significance of the correlations was determined by multiplying their p-value with the total number of edges in the graph, which was $717 * 716/2 = 256,686$, because we wanted to correct for multiple observations: we made as many observations as the number of edges in the graph. Thus, we obtained an upper limit of the p-value of the correlations. This meant that eight correlations have been deemed insignificant, but all the other edges of the spanning tree belonged to significant

correlations (this meant $716 - 8 = 708$ edges). This indicated that the attributes can be sorted into 9 clusters of tight dependency.

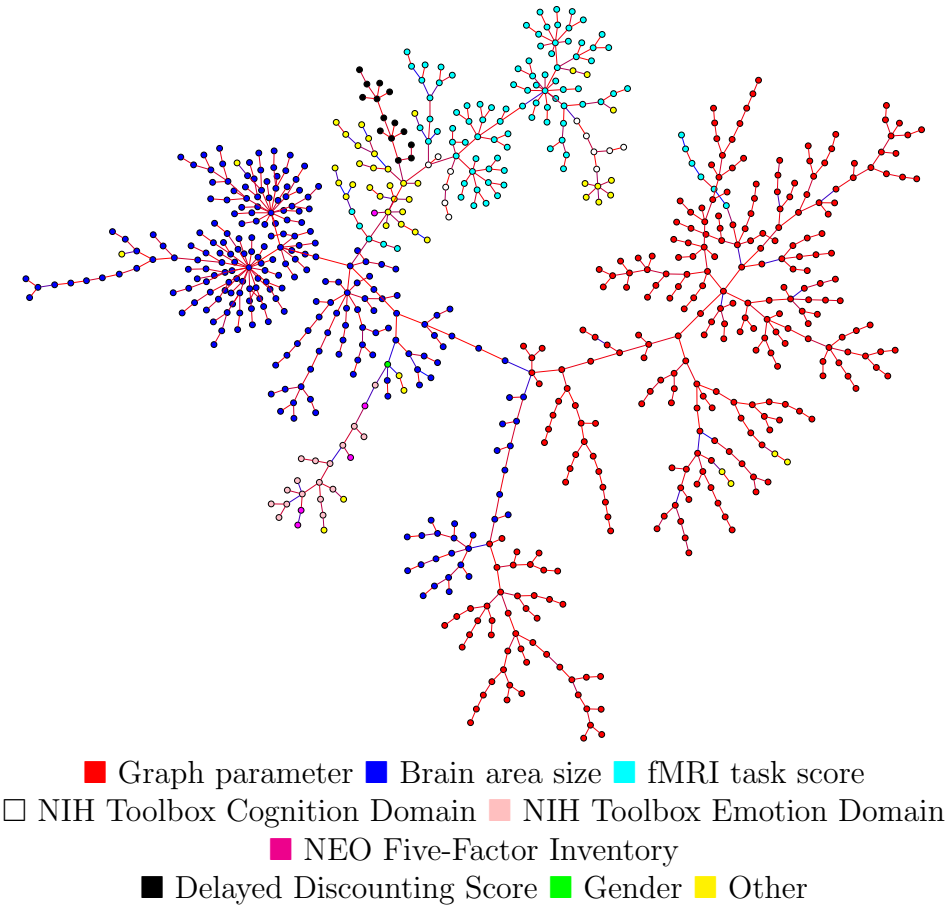

Figure 16: The maximum-weight spanning tree of the correlations of 717 quantities.

185 correlations referred to attributes which are either completely dependent on each other or are very close to a linear dependence (over 90% positive or negative correlation). Gray matter volume correlated with the volume of many cortical regions, which is not very surprising since these regions comprise the cortical gray matter. These included the superior frontal gyrus, the lateral orbitofrontal cortex, the precuneus, the middle temporal gyrus, the precentral gyrus, the fusiform gyrus and the rostral middle frontal gyrus (the list is incomplete).

Elements of the NIH Toolbox Emotion Domain showed a strong correlation with each other: sadness and fear affect, sadness and anger affect, sadness and perceived stress, sadness and loneliness, loneliness and perceived rejection, perceived hostility

and perceived rejection, friendship and emotional support, friendship and loneliness (negative), life satisfaction and meaning and purpose, emotional and instrumental support, life satisfaction and positive affect, anger-hostility and perceived stress, life satisfaction and perceived stress (negative), perceived stress and self-efficacy (negative), fear affect and fear-somatic arousal. Even the weakest of these correlations was 49% strong, the strongest (sadness and fear affect) being 72% strong.

There were 17 attributes in the NIH Toolbox Emotion Domain, and they almost represented a connected subgraph in the spanning tree (see Figure 17). This means that, by including the NEO-FFI Agreeableness attribute (which correlates positively with NIH Emotional Support and negatively with NIH Anger-Physical Aggression), we get an 18-vertex set which spans a 17-edge subtree of the spanning tree. This means that these attributes are strongly correlated with each other, comprising an important correlated subset of all the attributes.

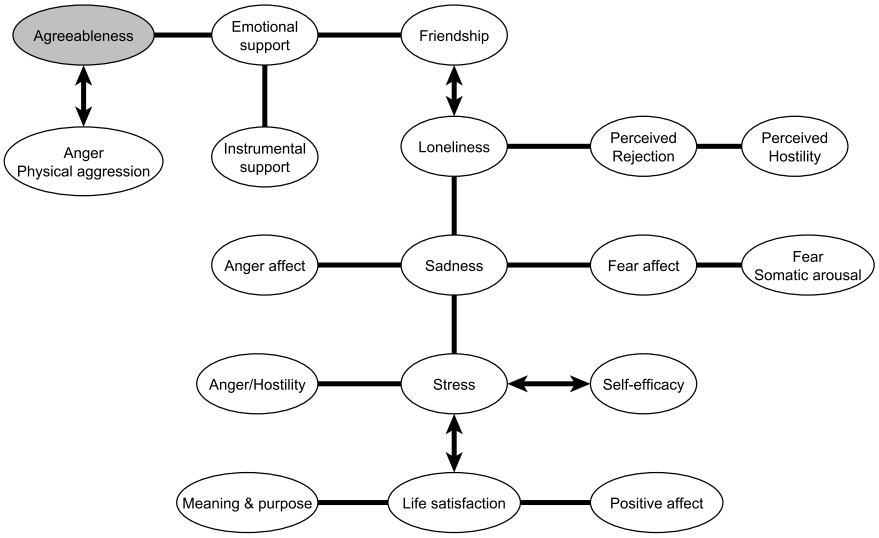

Figure 17: The attributes of the NIH Toolbox Emotion Domain. The connections are significant correlations in the constructed spanning tree.

The NEO-FFI personality scores appear to be correlated with certain NIH toolbox items, in a sense that they are leaves of those NIH toolbox items in the tree. NEO-FFI Neuroticism is a leaf of NIH Perceived Stress (correlation: 70%). NEO-FFI Conscientiousness is a leaf of NEO-FFI Neuroticism (correlation: -41%). NEO-FFI Extraversion is a leaf of NIH Friendship (correlation: 50%). An exception is

NEO-FFI Agreeableness, which is not a leaf as it is connected to both NIH Anger-Physical Aggression (correlation: -45%) and NIH Emotional Support (correlation: 35%). An interesting finding is that NEO-FFI Openness to Experience is a leaf of the NIH Toolbox Oral Reading Recognition Test Unadjusted Scale Score (correlation: 32%). This indicates that good reading skills and openness to experience frequently come together. To sum up, we can observe a strong connection between the NEO-FFI personality scores and the NIH toolbox positive or negative affect scores. This can mean multiple things. Our interpretation is that our personality defines our emotions to a great extent, which can be measured by statistical means. It should also be emphasized that this subtree may be suitable for a tree-based clustering of human emotions.

The NIH Toolbox Oral Reading Recognition Test is a very important hub in the spanning tree. Its unadjusted and age-adjusted versions are correlated with: Picture vocabulary task age-adjusted score (71%), Language task accuracy (49%), Penn Progressive Matrices Number of Correct Responses (47%), NIH Toolbox 2-minute Walk Endurance Test Age-Adjusted Score (43%), Delay Discounting Subjective Value for $40K at 1 year (35%), Short Penn Continuous Performance Test Specificity (33%), NEO-FFI Openness to Experience (32%), MMSE score (31%) and Penn Word Memory Test Total Number of Correct Responses (31%). Most of these are cognitive tests. The walk endurance test score is very interesting since it draws a connection between physical and intellectual fitness. It also seems that people who score higher on the oral reading test value a delayed payment higher than those with lower scores. There can be multiple reasons for this; the observed correlation could follow from their possible better financial status, financial skills or greater patience. We have covered the NEO-FFI connection above.

It seems that gender (male=1, female=2) correlates with the NIH Toolbox Grip Strength Test Age-Adjusted Scale Score (-75%), the brain mask volume (-67%), the optic chiasm volume (-41%) and the NIH Toolbox Anger-Physical Aggression Survey Unadjusted Scale Score (-26%). All of these are significant. This means that men have greater grip strength, brain volume (especially optic chiasm volume) and are

more aggressive physically.

The score of the walk endurance test correlates with the taste intensity score, although with one of the smallest significant observed correlations (-25%). Walk endurance seems to correlate with less perceived bitterness of quinine. The cause of this correlation is unknown to us, but it may mean that people with high walk endurance tolerate everything better, including bitter taste.

Another interesting correlation is NIH Toolbox 9-hole Pegboard Dexterity Test: Age-Adjusted Scale Score, versus Maximum matching (left hemisphere, weight function: mean fractional anisotropy) (27%). The corrected p-value is rather high when compared to the other correlations (p=0.003) but still significant. This is an important correlation because it shows a significant relationship between a parameter of the brain graph (connectome) and a brain performance score.

We expected the Age attribute to be a major hub of the spanning tree, but, interestingly, it was only a leaf of the attribute Right hemisphere cortical gray matter relative volume (correlation: -23%, p=0.037). We think this is because all the subjects were rather young, so their cognitive and psychological scores did not depend heavily on their age. Still, it is interesting that cortical gray matter relative volume correlates significantly with age, even among relatively young subjects.

7.4.2 Maximum spanning tree of Spearman's rank correlations

We also investigated how the spanning tree changes if we use Spearman's rank correlation coefficient instead of the classical (Pearson's) correlation coefficient [127]. We computed the rank correlation of two attributes as follows: the values were ordered independently for the two attributes, and then each value was substituted with its rank in the ordering. If two or more attributes were equal, then their rank was defined as the same number, and the succeeding rank(s) were omitted. In other words, the rank of a value v was defined as the number of values greater than v, plus one. For example, the values $1, 1, 4, 13, 13, 45$ were assigned the ranks $1, 1, 3, 4, 4, 6$.

We calculated a maximum weight spanning tree for the rank correlations as

well. The interactive figure with vertex labels is available at the address `http://pitgroup.org/static/graphmlviewer/index.html?src=correl_spanning_tree_rank.graphml`

This tree will be referred to as the *rank spanning tree* from now on, and the one using the traditional correlation coefficient will be referred to as the *original spanning tree*.

We examined those edges which are present in exactly one of the spanning trees. In the following analysis, we omitted the edges concerning two graph parameters or two brain area sizes. We also omitted those that connect two nodes (attributes) that can be exactly calculated from each other (e.g. the number of false positives and true negatives for some test). Edges which connect two attributes referring to subscores of the same task are also omitted.

There were 98 edges in each spanning tree that were not present in the other tree. As the trees contained 716 edges each, this means that approximately 13.7% of the edges were unique to the containing tree. In other words, the two trees were rather similar, having 86.3% of the edges in common. Among the unique edges, the vast majority were omitted from the analysis, or connected very similar nodes. For example, the unadjusted and age-adjusted version of an attribute, or the median reaction time and the number of correct responses for a task, or scores for two closely related tasks were considered similar attributes.

Some cognitive attributes were connected in a different way in the rank correlation tree when compared to the original correlation tree. For example, the IWRD and MMSE total scores were connected to the English reading score in the original tree, but connected to the Picture Vocabulary score in the rank tree. The computed score for the NIH Toolbox Words-In-Noise test was connected to a sub-score of the Social fMRI task in the original tree, but connected to, again, the Picture Vocabulary score in the rank tree. This may suggest that the Picture Vocabulary score is strongly connected to these complex cognitive scores, but not necessarily in a linear fashion. The same goes for the Working Memory fMRI task, which was connected to the Picture Vocabulary score in the original tree, but connected to both a sub-score

of the Relational task and the number of correct responses in the Penn Matrix Test in the rank tree.

An interesting connection between the volume of the right lateral ventricle and the number of correct happy identifications in the Penn Emotion Recognition Test was included in the rank correlation tree. However, we should note that the corrected p-value for this edge was very large, about 72. This means that there is a large likelihood that this connection was included by mere chance.

Regarding the graph parameters and the brain ROI sizes, the volume of the anterior corpus callosum and the relative volume of the mid-posterior corpus callosum were connected to two graph parameters (sum and minimum cut) in the original tree, but, surprisingly, these natural connections were no longer present in the rank tree. A graph parameter related to eigenvalues (Graph_Left_AdjLMaxDivD_FiberNDivLength) was connected to the NIH Toolbox Odor Identification Test unadjusted score in the normal tree, but this connection was not included in the rank tree, which, on the other hand, contained an edge between Age and the age-adjusted score of the odor identification test (corrected p-value: 6%. That connection is somewhat surprising since the age-adjusted scores should not be correlated with age. This could mean that some tests are not well adjusted for age. Another possible explanation could be that, although the test score does not change significantly for the same person over the person's lifetime, but different generations may have different mean scores due to environmental factors.

To sum up, we have analyzed both the original Pearson's and Spearman's rank correlations with 717 psychological, anatomical and connectome-properties originated from the Human Connectome Project's subject 500-release. Apart from numerous natural correlations, we have discovered numerous new significant correlations in the dataset. Brain graphs computed by us from the HCP data are available at `http://braingraph.org/download-pit-group-connectomes/`.

8 Appendix

8.1 Association Rule Mining and Alzheimer's Disease: Tables

Table 6: Basic statistics on the subjects of the CAMD data

Age distribution		**Gender distribution**		**MMSE distribution**	
A: up to 65 years	1093	Female	3315	A: severe cog. impairment	255
B: 66-75 years	2070	Male	2653	B: moderate cog. impairment	611
C: 76-85 years	2408			C: mild cog. impairment	3224
D: more than 85	397			D: normal cognition	1352

Table 8: Several association rules of the highest lift. The lift value describes the multiplication factor, increasing the probability of the Right Hand Side (RHS) if the Left Hand Side is true. For example, our best rule (the first below) is saying that one can have the a bad result of a cognitive test with four times higher probability if one has high serum sodium and either low cholesterol or low or normal blood glucose level.

```
(lb_sodium=h) & (lb_chol=l or lb_gluc=ln) ---> mm_ori=B
Universe: 2783, LHS support: 87, RHS support: 401, Support: 50
Confidence: 0.574713, Lift: 3.98859, Leverage: 0.0134618, p-value: 0, E-value: 0
3.98859

(lb_gluc=ln) & (lb_chol=l or lb_sodium=h) ---> mm_ori=B
Universe: 2783, LHS support: 105, RHS support: 401, Support: 57
Confidence: 0.542857, Lift: 3.76751, Leverage: 0.0150451, p-value: 0, E-value: 0
3.76751

(lb_sodium=h) & (lb_hct=l or lb_gluc=ln) ---> mm_ori=B
Universe: 2926, LHS support: 95, RHS support: 420, Support: 51
Confidence: 0.536842, Lift: 3.74, Leverage: 0.0127695, p-value: 0, E-value: 0
3.74

(lb_sodium=h) & (bpsys=ln or lb_gluc=ln) ---> mm_ori=B
Universe: 3091, LHS support: 102, RHS support: 425, Support: 52
Confidence: 0.509804, Lift: 3.70777, Leverage: 0.0122858, p-value: 0, E-value: 0
3.70777

(lb_gluc=ln) & (lb_creat=l or lb_sodium=h) ---> mm_ori=B
Universe: 3091, LHS support: 99, RHS support: 425, Support: 50
```

Confidence: 0.505051, Lift: 3.6732, Leverage: 0.0117722, p-value: 0, E-value: 0
3.6732

(lb_sodium=h) & (age=D or lb_gluc=ln) ---> mm_ori=B
Universe: 3091, LHS support: 101, RHS support: 425, Support: 51
Confidence: 0.50495, Lift: 3.67248, Leverage: 0.0120068, p-value: 0, E-value: 0
3.67248

(lb_gluc=ln) & (lb_ast=l or lb_sodium=h) ---> mm_ori=B
Universe: 3091, LHS support: 101, RHS support: 425, Support: 51
Confidence: 0.50495, Lift: 3.67248, Leverage: 0.0120068, p-value: 0, E-value: 0
3.67248

Table 9: Some association rules involving serum cholesterol level

(lb_sodium=h) & (lb_chol=l or lb_gluc=ln) ---> mm_ori=B
Universe: 2783, LHS support: 87, RHS support: 401, Support: 50
Confidence: 0.574713, Lift: 3.98859, Leverage: 0.0134618, p-value: 0, E-value: 0
3.98859

(lb_gluc=ln) & (lb_chol=l or lb_sodium=h) ---> mm_ori=B
Universe: 2783, LHS support: 105, RHS support: 401, Support: 57
Confidence: 0.542857, Lift: 3.76751, Leverage: 0.0150451, p-value: 0, E-value: 0
3.76751

(lb_sodium=h) & (lb_chol=ln or lb_gluc=ln) ---> mm_ori=B
Universe: 2783, LHS support: 106, RHS support: 401, Support: 55
Confidence: 0.518868, Lift: 3.60102, Leverage: 0.0142747, p-value: 0, E-value: 0
3.60102

(lb_chol=h) & (lb_cl=h or lb_sodium=h) ---> mm_ori=B
Universe: 1420, LHS support: 71, RHS support: 304, Support: 51
Confidence: 0.71831, Lift: 3.35526, Leverage: 0.0252113, p-value: 2.22045e-016, E-value: 1.88773e-007
3.35526

(lb_chol=h) & (lb_monole=l or lb_sodium=h) ---> mm_total=AB
Universe: 1364, LHS support: 73, RHS support: 325, Support: 58
Confidence: 0.794521, Lift: 3.33454, Leverage: 0.02977, p-value: 1.51101e-013, E-value: 0.00012846
3.33454

(lb_sodium=h) & (lb_monole=h or lb_chol=h) ---> mm_total=AB
Universe: 1364, LHS support: 66, RHS support: 325, Support: 51
Confidence: 0.772727, Lift: 3.24308, Leverage: 0.0258608, p-value: 5.9952e-015, E-value: 5.09687e-006
3.24308

(lb_chol=h) & (lb_hbsag=h or lb_sodium=h) ---> mm_attcal=B

```
Universe: 1164, LHS support: 67, RHS support: 312, Support: 50
Confidence: 0.746269, Lift: 2.78416, Leverage: 0.0275268, p-value: 6.2725e-011, E-value: 0.0533262
2.78416

(lb_sodium=h) & (lb_bun=h or lb_chol=h) ---> mm_attcal=B
Universe: 1387, LHS support: 61, RHS support: 429, Support: 52
Confidence: 0.852459, Lift: 2.75609, Leverage: 0.023888, p-value: 8.87168e-012, E-value: 0.00754232
2.75609

(lb_sodium=h) & (lb_ca=l or lb_chol=h) ---> mm_attcal=B
Universe: 1420, LHS support: 61, RHS support: 460, Support: 51
Confidence: 0.836066, Lift: 2.5809, Leverage: 0.0219996, p-value: 2.52266e-011, E-value: 0.0214466
2.5809

(lb_sodium=h) & (lb_cl=h or lb_chol=h) ---> mm_attcal=B
Universe: 1420, LHS support: 66, RHS support: 460, Support: 55
Confidence: 0.833333, Lift: 2.57246, Leverage: 0.0236759, p-value: 1.65421e-010, E-value: 0.140634
2.57246
```

Table 10: Legends for Table 8 and 9

age	Subject age (A: max. 65 years, B: 66–75 years, C: 76–85 years, D: >85 years)
ast_alt	De Ritis ratio
bpdia	Diastolic blood pressure
bpsys	Systolic blood pressure
lb_alb	Serum albumine
lb_alp	Serum alkaline phosphatase
lb_alt	Serum alanine aminotransferase
lb_ast	Serum aspartate aminotransferase
lb_baso	Basophils, particle concentration
lb_bili	Serum indirect bilirubin
lb_bun	Blood Urea Nitrogen
lb_ca	Serum calcium
lb_chol	Serum cholesterol
lb_ck	Serum creatine kinase
lb_cl	Serum chlorine
lb_creat	Serum creatinine
lb_eos	Eosinophils, particle concentration
lb_gluc	Serum glucose
lb_hba1c	Hemoglobin A1C
lb_hbsag	Hepatitis B virus surface antigen
lb_hct	Hematocrit
lb_hgb_blood	Blood hemoglobin
lb_k	Serum potassium
lb_ketones	Ketones

lb_ldh	Lactate dehydrogenase
lb_lym	Lymphocytes, particle concentration
lb_lymle	Lymphocytes/leukocytes ratio
lb_mch	Mean Corpuscular Hemoglobin
lb_mchc	Mean Corpuscular Hemoglobin Concentration
lb_mcv	Mean Corpuscular Volume
lb_mono	Monocytes, particle concentration
lb_monole	Monocytes/leukocytes ratio
lb_neut	Neutrophils, particle concentration
lb_neutle	Neutrophils/leukocytes ratio
lb_ph	pH
lb_phos	Phosphate
lb_plat	Platelets
lb_prot	Total protein
lb_rbc_blood	Red blood count
lb_sodium	Serum sodium
lb_tsh	Thyrotropin
lb_vitb12	Serum B12 vitamin
lb_wbc_blood	White blood count
mm_attcal	MMSE attention and calculation subscore (B: 0–1, C: 2, D: 3, E: 4–5)
mm_lang	MMSE language subscore (B: 0–2, C: 3–4, D: 5–6, E: 7–9)
mm_ori	MMSE orientation subscore (B: 0–2, C: 3–4, D: 5–7, E: 8–10)
mm_recall	MMSE recall subscore (B: 0, C: 1, D: 2, E: 3)
mm_total	MMSE total score (A: <10, B: 10–14, C: 15–23, D: $\geq$ 24)
pulse	Heart rate
resp	Respiratory rate
sex	Subject sex (F: female, M: male)
temper	Temperature

Table 12: Elementary clauses with the greatest positive effect on normal cognition

lb_vitb12=h	score: 67
lb_mch=h	score: 25
lb_mchc=l	score: 22
lb_k=h	score: 17
sex=M	score: 10
pulse=l	score: 9
lb_bun=l	score: 8
age=AB	score: 4
lb_mono=nh	score: 3
resp=ln	score: 3
lb_plat=ln	score: 2
lb_eos=nh	score: 2
lb_prot=nh	score: 2

Table 14: Elementary clauses with the greatest negative effect on normal cognition

temper=nh	score: -10
lb_wbc_blood=h	score: -10
age=BCD	score: -10
lb_prot=h	score: -12
lb_gluc=h	score: -12
pulse=h	score: -12
lb_ck=h	score: -12
lb_hct=nh	score: -12
lb_k=ln	score: -12
lb_alp=h	score: -12
lb_chol=ln	score: -13
lb_ph=h	score: -13
lb_hct=l	score: -13
lb_alt=h	score: -13
age=A	score: -14
bpsys=ln	score: -14
lb_creat=ln	score: -14
lb_creat=h	score: -16
temper=l	score: -17
lb_alp=ln	score: -18
lb_bun=ln	score: -18
lb_alt=l	score: -19
lb_wbc_blood=l	score: -20
lb_chol=l	score: -21
pulse=nh	score: -21
lb_prot=ln	score: -22
lb_bun=h	score: -22
lb_plat=h	score: -26
lb_gluc=ln	score: -27
bpdia=ln	score: -28
age=CD	score: -32
lb_chol=h	score: -42
lb_ast=h	score: -43
lb_ca=l	score: -50
sex=F	score: -57
age=D	score: -99
lb_cl=h	score: -173
lb_sodium=h	score: -224

8.2 Differences between female and male connectomes

The following table contains the results and the statistical analysis of the graph-theoretical evaluation of the sex differences in the 96 diffusion MRI images. The first column gives the resolution in each hemisphere; the number of nodes in the whole graph is 83, 129 and 234, respectively. The second column describes the graph parameter computed: its syntactics is as follows: each parameter-name contains two separating "_" symbols that define three parts of the parameter-name. The first part describe the hemisphere or the whole connectome with the words Left, Right or All. The second part describes the parameter computed, and the third part the weight function used. The third column contains the p-values of the first round, the fourth column the p-values of the second round, and the fifth column the (very strict) Holm-Bonferroni correction of the p-value. With p=0.05, the first 12 rows *all* describe significantly different graph properties between sexes. On the other hand, each row with an italic fourth column describes *in itself* a significant difference between sexes, again with p=0.05, but together these may not be significant.

Scale	Property	p (1st)	p (2nd)	p (corrected)
129	Right_MinCutBalDivSum_FAMean	0.00807	*0.00003*	**0.00401**
83	All_LogSpanningForestN_FiberNDivLength	0.00003	*0.00004*	**0.00451**
234	All_PGEigengap_FiberNDivLength	0.00321	*0.00007*	**0.00798**
129	All_PGEigengap_FiberNDivLength	0.00792	*0.00011*	**0.01303**
83	Left_MinCutBalDivSum_FiberN	0.00403	*0.00011*	**0.01300**
83	Right_MinCutBalDivSum_FAMean	0.00496	*0.00015*	**0.01744**
129	Left_PGEigengap_FiberNDivLength	0.00223	*0.00015*	**0.01797**
234	All_PGEigengap_FiberN	0.00826	*0.00022*	**0.02517**
83	All_Sum_Unweighted	0.00025	*0.00022*	**0.02504**
129	Left_MinCutBalDivSum_FiberN	0.00001	*0.00023*	**0.02563**
83	All_LogSpanningForestN_FiberN	0.00001	*0.00028*	**0.03084**
83	Right_Sum_FAMean	0.00028	*0.00029*	**0.03224**
234	All_Sum_Unweighted	0.00063	*0.00032*	**0.03512**
234	Left_PGEigengap_FiberNDivLength	0.00013	*0.00038*	**0.04171**

129	All_Sum_Unweighted	0.00026	*0.00042*	**0.04563**
234	All_Sum_FAMean	0.00014	*0.00047*	**0.04988**
129	All_LogSpanningForestN_FiberN	0.00000	*0.00048*	0.05045
83	All_Sum_FAMean	0.00029	*0.00050*	0.05260
129	Right_Sum_FAMean	0.00062	*0.00051*	0.05355
234	Right_PGEigengap_FiberNDivLength	0.00041	*0.00053*	0.05414
83	Left_Sum_Unweighted	0.00378	*0.00068*	0.06936
234	Right_Sum_FAMean	0.00085	*0.00084*	0.08454
234	Left_Sum_Unweighted	0.00293	*0.00092*	0.09212
129	All_Sum_FAMean	0.00015	*0.00097*	0.09650
234	Left_MinCutBalDivSum_FiberN	0.00002	*0.00108*	0.10539
83	Left_LogSpanningForestN_FiberNDivLength	0.00343	*0.00116*	0.11274
83	All_LogSpanningForestN_Unweighted	0.00113	*0.00121*	0.11629
234	Left_MinCutBalDivSum_FiberLengthMean	0.00411	*0.00123*	0.11646
83	All_LogSpanningForestN_FAMean	0.00012	*0.00126*	0.11823
83	Right_Sum_Unweighted	0.00019	*0.00128*	0.11891
129	Left_MinCutBalDivSum_Unweighted	0.00265	*0.00134*	0.12351
83	Left_MinCutBalDivSum_Unweighted	0.00206	*0.00136*	0.12370
129	Left_PGEigengap_FiberN	0.00382	*0.00142*	0.12775
234	All_LogSpanningForestN_FAMean	0.00043	*0.00150*	0.13343
234	Left_PGEigengap_FiberN	0.00066	*0.00163*	0.14369
129	Right_LogSpanningForestN_FAMean	0.00143	*0.00170*	0.14769
83	Left_MinCutBalDivSum_FiberNDivLength	0.00031	*0.00175*	0.15023
129	All_LogSpanningForestN_FiberNDivLength	0.00000	*0.00177*	0.15009
129	All_LogSpanningForestN_Unweighted	0.00218	*0.00182*	0.15279
129	Right_Sum_Unweighted	0.00068	*0.00186*	0.15417
129	Left_PGEigengap_FAMean	0.00995	*0.00191*	0.15694
129	All_LogSpanningForestN_FAMean	0.00019	*0.00211*	0.17093
234	Left_Sum_FAMean	0.00026	*0.00212*	0.16978
83	Right_LogSpanningForestN_FAMean	0.00067	*0.00239*	0.18842
234	Left_PGEigengap_FAMean	0.00141	*0.00240*	0.18684
83	Left_PGEigengap_Unweighted	0.00458	*0.00243*	0.18738

129	Left_MinCutBalDivSum_FiberLengthMean	0.00892	*0.00245*	0.18596
83	Left_Sum_FAMean	0.00056	*0.00279*	0.20893
234	Left_MinCutBalDivSum_Unweighted	0.00154	*0.00289*	0.21355
234	Left_PGEigengap_FiberLengthMean	0.00554	*0.00295*	0.21516
234	Right_LogSpanningForestN_FAMean	0.00380	*0.00305*	0.21935
234	Left_PGEigengap_Unweighted	0.00176	*0.00338*	0.24029
83	Left_PGEigengap_FAMean	0.00215	*0.00359*	0.25152
83	Left_LogSpanningForestN_FiberN	0.00012	*0.00395*	0.27269
129	Left_Sum_Unweighted	0.00232	*0.00456*	0.31006
83	Left_LogSpanningForestN_FAMean	0.00082	*0.00496*	0.33212
234	Right_MinCutBalDivSum_Unweighted	0.00462	*0.00543*	0.35825
83	Right_LogSpanningForestN_FiberNDivLength	0.00022	*0.00587*	0.38180
234	Left_LogSpanningForestN_FAMean	0.000129	*0.00595*	0.38054
234	Right_PGEigengap_Unweighted	0.00095	*0.00626*	0.39459
129	Left_Sum_FAMean	0.00032	*0.00660*	0.40907
83	Left_AdjLMaxDivD_FiberN	0.00501	*0.00804*	0.49040
234	Right_Sum_Unweighted	0.00224	*0.00845*	0.50692
234	Right_PGEigengap_FiberN	0.00009	*0.00910*	0.53671
129	All_Sum_FiberN	0.00000	*0.00938*	0.54418
234	Right_PGEigengap_FAMean	0.00074	*0.00974*	0.55538
129	Right_PGEigengap_FAMean	0.00296	*0.00981*	0.54933
83	Right_PGEigengap_Unweighted	0.00087	*0.01053*	0.57889
129	Right_MinCutBalDivSum_FiberN	0.00563	*0.01101*	0.59432
129	Right_MinCutBalDivSum_Unweighted	0.00492	*0.01212*	0.64227
129	Left_LogSpanningForestN_FAMean	0.00106	*0.01218*	0.63359
129	Left_LogSpanningForestN_FiberN	0.00014	*0.01258*	0.64134
83	All_Sum_FiberN	0.00000	*0.01290*	0.64480
234	All_Sum_FiberN	0.00000	*0.01358*	0.66520
83	Right_LogSpanningForestN_Unweighted	0.00541	*0.01438*	0.69010
129	Left_LogSpanningForestN_FiberNDivLength	0.00288	*0.01447*	0.67995
129	Right_PGEigengap_Unweighted	0.00242	*0.01676*	0.77084
129	Right_PGEigengap_FiberN	0.00869	*0.01706*	0.76750

234	All_MinVertexCover_FAMean	0.00289	*0.01713*	0.75373
83	All_HoffmanBound_FAMean	0.00087	*0.02011*	0.86462
83	All_Sum_FiberNDivLength	0.00002	*0.02117*	0.88929
234	Right_MinCutBalDivSum_FiberN	0.00234	*0.02197*	0.90065
83	Right_LogSpanningForestN_FiberN	0.00083	*0.02539*	1.01567
234	Right_MinCutBalDivSum_FiberLengthMean	0.00234	*0.02663*	1.03841
83	Right_MinCutBalDivSum_FiberNDivLength	0.00072	*0.02854*	1.08446
129	Left_MinCutBalDivSum_FiberNDivLength	0.00019	*0.02897*	1.07195
83	Right_PGEigengap_FAMean	0.00112	*0.02948*	1.06119
234	All_LogSpanningForestN_FiberN	0.00091	*0.03308*	1.15795
234	Right_PGEigengap_FiberLengthMean	0.00367	*0.03369*	1.14542
129	Right_MinCutBalDivSum_FiberLengthMean	0.00768	*0.04500*	1.48511
129	All_Sum_FiberNDivLength	0.00008	*0.04728*	1.51293
129	Right_LogSpanningForestN_FiberNDivLength	0.00051	*0.04891*	1.51627
234	All_LogSpanningForestN_FiberNDivLength	0.00106	0.05095	1.52842
129	Right_LogSpanningForestN_FiberN	0.00045	0.05578	1.61751
83	Right_MinCutBalDivSum_FiberN	0.00346	0.06284	1.75951
83	Right_HoffmanBound_FiberNDivLength	0.005129	0.06309	1.70341
83	Right_PGEigengap_FiberLengthMean	0.00949	0.06515	1.69395
234	Left_MinCutBalDivSum_FiberNDivLength	0.00642	0.06548	1.63696
234	Left_MinVertexCover_FAMean	0.00107	0.07139	1.71336
234	All_Sum_FiberNDivLength	0.00044	0.07318	1.68305
83	Right_Sum_FiberN	0.00000	0.07799	1.71586
83	Right_Sum_FiberNDivLength	0.00018	0.07920	1.66329
129	Left_Sum_FiberN	0.00000	0.08380	1.67598
129	Right_Sum_FiberN	0.00001	0.08653	1.64406
129	Left_HoffmanBound_Unweighted	0.00848	0.08944	1.60984
83	Left_Sum_FiberN	0.00000	0.09430	1.60310
234	Left_Sum_FiberN	0.00040	0.11447	1.83157
129	Right_Sum_FiberNDivLength	0.00180	0.12102	1.81523
234	Right_Sum_FiberN	0.00012	0.16411	2.29752
83	Left_Sum_FiberNDivLength	0.00043	0.16774	2.18062

129	Left_Sum_FiberNDivLength	0.00100	0.22542	2.70502
234	Right_Sum_FiberNDivLength	0.00562	0.23691	2.60604
83	Right_HoffmanBound_FAMean	0.00587	0.32069	3.20692
83	All_MinVertexCoverBinary_Unweighted	0.00716	0.38829	3.49459
234	Right_LogSpanningForestN_FiberNDivLength	0.00940	0.40996	3.27971
83	Left_HoffmanBound_FiberN	0.00175	0.41913	2.93394
83	All_MinVertexCover_FiberNDivLength	0.00036	0.46677	2.80065
83	Right_MinSpanningForest_FiberLengthMean	0.00491	0.55239	2.76195
234	Right_MinSpanningForest_FiberLengthMean	0.00601	0.55631	2.22523
129	All_MinVertexCover_FiberN	0.00232	0.71406	2.14217
83	All_MinVertexCover_FiberN	0.00244	0.84437	1.68874
234	All_MinVertexCover_FiberN	0.00055	0.92958	0.92958

9 One-page summary (English)

Biological research often results in big datasets. Without efficient computer algorithms, analysis of these large datasets would certainly be impossible.

The present work demonstrates the application of mathematics and informatics in multiple areas of biology. The first section generalizes the well-known k-means algorithm for non-Euclidean data points. This generalization has since been utilized by multiple other researchers.

An application of association rule mining in Alzheimer's disease is also presented. Data for more than 6000 subjects in 11 drug trials has been analyzed to find connections between combinations of biomarkers and dementia. The results show that, in certain age groups, even simple combinations of biomarkers can be accurate predictors of low cognitive scores.

By analyzing the gut metagenomes of diabetic and non-diabetic subjects it was demonstrated that short DNA sequences of length at most 9 can be associated with diabetes. This may be the first time that the frequency of sequences of this length is associated with a medical condition.

Next-generation sequencing methods allow us to obtain information about microorganisms without culturing. We can also use it to find homologous genes in a metagenome sequenced from an environmental sample. A method called "the metagenomical telescope" is described, which enriches a Hidden Markov Model of a gene family with genes found in an environmental sample, then uses the enriched model on genomes of model organisms to characterize proteins with sofar unknown functions.

The human brain is a network of smaller areas, also referred to as ROIs (regions of interest). The graph of ROIs connected by nerve fibers can be constructed by tracing MRI imagery. Graphs of this kind were analyzed by graph theoretical methods. Significant differences have been found between the brain graphs of women and men. A web application (Budapest Reference Connectome) with an interactive 3D brain graph was also developed.

The Human Connectome Project provided supplementary data such as cognitive, emotional and personality scores and demographics. We calculated correlations between these attributes, interpreted these as a weighted graph, then constructed a maximum spanning tree to highlight the most important connections. This is the first time this method has been applied to a biological dataset. Applications of this spanning tree include a reasonable clustering of emotional states.

10 One-page summary (Hungarian)

A biológiai kutatások gyakran nagyon sok adatot eredményeznek. Hatékony számítógépes algoritmusok nélkül nem lenne lehetséges ezeknek az adathalmazoknak az elemzése.
Ez a munka a matematika és informatika alkalmazásait demonstrálja a biológia több területén. Az első fejezet az ismert k-means algoritmust általánosítja nem-euklideszi adatpontokra. Ezt az általánosítást azóta több más kutató is felhasználta.
Az asszociációsszabály-bányászat egy alkalmazása az Alzheimer-kórral kapcsolatban is bemutatásra kerül. Több mint 6000 alany 11 gyógyszerkísérletből származó adatait elemeztük, hogy kapcsolatot találjunk biomarkerek kombinációi és a demencia közt. Az eredmények azt mutatják, hogy bizonyos korcsoportokban még az egyszerű kombinációk is pontosan jelezhetik az alacsony pontszámokat kognitív teszteken.
Diabéteszes és nem diabéteszes alanyok bélflórájának metagenomját elemezve megmutattuk, hogy a max. 9 hosszú DNS szakaszok gyakorisága társítható a diabéteszhez. Ez lehet az egyik első alkalom, hogy ilyen hosszúságú szekvenciák kapcsolatba lettek hozva egy betegséggel.
Az új generációs szekvenálási módszerek lehetővé teszik, hogy tenyésztés nélkül nyerjünk adatot mikroorganizmusokról. Ezzel a módszerrel homológ géneket is találhatunk egy metagenomban. A "metagenomikai teleszkóp" nevű eljárásról szóló fejezetben egy rejtett Markov-modellt "dúsítunk fel" metagenomokból származó génekkel, hogy aztán ismeretlen funkciójú géneket jellemezzünk modellorganizmusok genomjában.
Az emberi agy kisebb területek, ROI-k (regions of interest) hálózata. Az idegrostok által összekötött ROI-k gráfja MR-felvételek feldolgozásával állítható elő. Az ilyen típusú gráfokat gráfelméleti módszerekkel elemeztük. Szignifikáns különbségeket találtunk a nők és férfiak agygráfja között. Egy interaktív 3D agygráfot megjelenítő webalkalmazást (Budapest Reference Connectome) is kifejlesztettünk.
A Human Connectome Project MR-felvételek mellett más adatot is közölt, pl. kogníciós és érzelmi tesztek eredményét, valamint demográfiai adatokat. Ezekből korrelációkat számoltunk, amelyekből a legfontosabbakat egy maximális feszítőfa módszerrel próbáltuk kiemelni. Ezt a módszert mi használtuk először biológiai adatokra. A feszítőfánk alkalmazásai közé tartozik például az érzelmi állapotok statisztikailag megalapozott klaszterezése.

References

[1] T. Achterberg, *SCIP: solving constraint integer programs*, Mathematical Programming Computation, 1 (2009), pp. 1–41.

[2] T. Achterberg, T. Berthold, T. Koch, and K. Wolter, *Constraint integer programming: A new approach to integrate CP and MIP*, in Integration of AI and OR techniques in constraint programming for combinatorial optimization problems, Springer, 2008, pp. 6–20.

[3] P. Ade, N. Aghanim, C. Armitage-Caplan, M. Arnaud, M. Ashdown, F. Atrio-Barandela, J. Aumont, C. Baccigalupi, A. Banday, R. Barreiro, et al., *Planck 2013 results. XVI. Cosmological parameters*, arXiv preprint arXiv:1303.5076, (2013).

[4] H. J. Adrogue and N. E. Madias, *Hypernatremia.*, N Engl J Med, 342 (2000), pp. 1493–1499.

[5] F. Agosta, S. Galantucci, P. Valsasina, E. Canu, A. Meani, A. Marcone, G. Magnani, A. Falini, G. Comi, and M. Filippi, *Disrupted brain connectome in semantic variant of primary progressive aphasia.*, Neurobiol Aging, (2014).

[6] R. Agrawal, T. Imielinski, and A. N. Swami, *Mining association rules between sets of items in large databases*, in Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data, Washington, D.C., May 26-28, 1993, P. Buneman and S. Jajodia, eds., ACM Press, 1993, pp. 207–216.

[7] A. F. Alexander-Bloch, P. T. Reiss, J. Rapoport, H. McAdams, J. N. Giedd, E. T. Bullmore, and N. Gogtay, *Abnormal cortical growth in schizophrenia targets normative modules of synchronized development.*, Biol Psychiatry, (2014).

[8] I. Alsmadi and I. Alhami, *Clustering and classification of email contents*, Journal of King Saud University - Computer and Information Sciences, 27 (2015), pp. 46 – 57.

[9] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman, *Basic local alignment search tool.*, J Mol Biol, 215 (1990), pp. 403–410.

[10] J. Amar, M. Serino, C. Lange, C. Chabo, J. Iacovoni, S. Mondot, P. Lepage, C. Klopp, J. Mariette, O. Bouchez, L. Perez, M. Courtney, M. Marre, P. Klopp, O. Lantieri, J. Dore, M. Charles, B. Balkau, R. Burcelin, and D. S. Group, *Involvement of tissue bacteria in the onset of diabetes in humans: evidence for a concept.*, Diabetologia, 54 (2011), pp. 3055–3061.

[11] J. R. Anderson, B. W. Jones, C. B. Watt, M. V. Shaw, J.-H. Yang, D. Demill, J. S. Lauritzen, Y. Lin, K. D. Rapp, D. Mastronarde, P. Koshevoy, B. Grimm, T. Tasdizen, R. Whitaker, and R. E. Marc, *Exploring the retinal connectome.*, Mol Vis, 17 (2011), pp. 355–379.

[12] M. Ankerst, M. M.Breunig, H. Kriegel, and J. Sander, *Optics: Ordering points to identify the clustering structure*, in Proc. ACM SIGMOD '99 Int. Conf. on Management of Data, Philadelphia PA, 1999.

[13] K. Arnold, L. Bordoli, J. Kopp, and T. Schwede, *The swiss-model workspace: a web-based environment for protein structure homology modelling.*, Bioinformatics, 22 (2006), pp. 195–201.

[14] M. Arumugam, J. Raes, E. Pelletier, D. Le Paslier, T. Yamada, D. R. Mende, G. R. Fernandes, J. Tap, T. Bruls, J.-M. Batto, M. Bertalan, N. Borruel, F. Casellas, L. Fernandez, L. Gautier, T. Hansen, M. Hattori, T. Hayashi, M. Kleerebezem, K. Kurokawa, M. Leclerc, F. Levenez, C. Manichanh, H. B. Nielsen, T. Nielsen, N. Pons, J. Poulain, J. Qin, T. Sicheritz-Ponten, S. Tims, D. Torrents, E. Ugarte, E. G. Zoetendal,

J. Wang, F. Guarner, O. Pedersen, W. M. de Vos, S. Brunak, J. Doré, M. I. T. C. , M. Antolín, F. Artiguenave, H. M. Blottiere, M. Almeida, C. Brechot, C. Cara, C. Chervaux, A. Cultrone, C. Delorme, G. Denariaz, R. Dervyn, K. U. Foerstner, C. Friss, M. van de Guchte, E. Guedon, F. Haimet, W. Huber, J. van Hylckama-Vlieg, A. Jamet, C. Juste, G. Kaci, J. Knol, O. Lakhdari, S. Layec, K. Le Roux, E. Maguin, A. Mérieux, R. Melo Minardi, C. M'rini, J. Muller, R. Oozeer, J. Parkhill, P. Renault, M. Rescigno, N. Sanchez, S. Sunagawa, A. Torrejon, K. Turner, G. Vandemeulebrouck, E. Varela, Y. Winogradsky, G. Zeller, J. Weissenbach, S. D. Ehrlich, and P. Bork, *Enterotypes of the human gut microbiome.*, Nature, 473 (2011), pp. 174–180.

[15] G. Astarita, K.-M. Jung, N. C. Berchtold, V. Q. Nguyen, D. L. Gillen, E. Head, C. W. Cotman, and D. Piomelli, *Deficient liver biosynthesis of docosahexaenoic acid correlates with cognitive impairment in Alzheimer's disease.*, PLoS One, 5 (2010), p. e12538.

[16] B. J. Baker, G. W. Tyson, R. I. Webb, J. Flanagan, P. Hugenholtz, E. E. Allen, and J. F. Banfield, *Lineages of acidophilic archaea revealed by community genomic analysis.*, Science, 314 (2006), pp. 1933–1935.

[17] J. T. Baker, A. J. Holmes, G. A. Masters, B. T. T. Yeo, F. Krienen, R. L. Buckner, and D. Öngür, *Disruption of cortical association networks in schizophrenia and psychotic bipolar disorder.*, JAMA Psychiatry, 71 (2014), pp. 109–118.

[18] G. Ball, P. Aljabar, S. Zebari, N. Tusor, T. Arichi, N. Merchant, E. C. Robinson, E. Ogundipe, D. Rueckert, A. D. Edwards, and S. J. Counsell, *Rich-club organization of the newborn human brain.*, Proc Natl Acad Sci U S A, 111 (2014), pp. 7456–7461.

[19] C. I. Bargmann, *Beyond the connectome: how neuromodulators shape neural circuits.*, Bioessays, 34 (2012), pp. 458–465.

[20] P. J. Basser and C. Pierpaoli, *Microstructural and physiological features of tissues elucidated by quantitative-diffusion-tensor mri.*, J Magn Reson, 213 (1996), pp. 560–570.

[21] D. Batalle, E. Muñoz-Moreno, F. Figueras, N. Bargallo, E. Eixarch, and E. Gratacos, *Normalization of similarity-based individual brain networks from gray matter MRI and its association with neurodevelopment in infants with intrauterine growth restriction.*, Neuroimage, 83 (2013), pp. 901–911.

[22] C. Bauckhage, *Numpy / scipy recipes for data science: k-medoids clustering*, 2009.

[23] L. E. Baum and T. Petrie, *Statistical inference for probabilistic functions of finite state markov chains*, Ann. Math. Statist., 37 (1966), pp. 1554–1563.

[24] F. Bermejo-Pareja, J. Benito-Leon, S. Vega, M. J. Medrano, G. C. Roman, and N. D. in Central Spain (NEDICES) Study Group, *Incidence and subtypes of dementia in three elderly populations of central spain.*, J Neurol Sci, 264 (2008), pp. 63–72.

[25] G. Bernardi and G. Bernardi, *Compositional constraints and genome evolution.*, J Mol Evol, 24 (1986), pp. 1–11.

[26] P. Besson, V. Dinkelacker, R. Valabregue, L. Thivard, X. Leclerc, M. Baulac, D. Sammler, O. Colliot, S. Lehéricy, S. Samson, and S. Dupont, *Structural connectivity differences in left and right temporal lobe epilepsy.*, Neuroimage, 100C (2014), pp. 135–144.

[27] B. P. Boerner and N. E. Sarvetnick, *Type 1 diabetes: role of intestinal microbiome in humans and mice.*, Ann N Y Acad Sci, 1243 (2011), pp. 103–118.

[28] L. Bonilha, T. Nesland, C. Rorden, P. Fillmore, R. P. Ratnayake, and J. Fridriksson, *Mapping remote subcortical ramifications of injury after ischemic strokes.*, Behav Neurol, 2014 (2014), p. 215380.

[29] I. Borg and P. J. F. Groenen, *Modern Multidimensional Scaling*, Springer Series in Statistics, 2005.

[30] J.-P. Bouchaud and M. Potters, *Theory of financial risk and derivative pricing: from statistical physics to risk management*, Cambridge university press, 2003.

[31] W. D. Bradley, C. Zwingelstein, and C. M. Rondinone, *The emerging role of the intestine in metabolic diseases.*, Arch Physiol Biochem, 117 (2011), pp. 165–176.

[32] J. G. Brida, M. Deidda, N. Garrido, and P. Manuela, *Analyzing the performance of the south tyrolean hospitality sector: a dynamic approach*, International Journal of Tourism Research, 17 (2015), pp. 196–208.

[33] R. Buffa, R. M. Mereu, P. F. Putzu, G. Floris, and E. Marini, *Bioelectrical impedance vector analysis detects low body cell mass and dehydration in patients with Alzheimer's disease.*, J Nutr Health Aging, 14 (2010), pp. 823–827.

[34] P. D. Cani and N. M. Delzenne, *The gut microbiome as therapeutic target.*, Pharmacol Ther, 130 (2011), pp. 202–212.

[35] A. D. Carlo, M. Baldereschi, L. Amaducci, V. Lepore, L. Bracco, S. Maggi, S. Bonaiuto, E. Perissinotto, G. Scarlato, G. Farchi, D. Inzitari, and I. L. S. A. W. Group, *Incidence of dementia, Alzheimer's disease, and vascular dementia in Italy. the ILSA study.*, J Am Geriatr Soc, 50 (2002), pp. 41–48.

[36] R. Chaves, J. M. Gorriz, J. Ramirez, I. A. Illan, D. Salas-Gonzalez, and M. Gomez-Rio, *Efficient mining of association rules for the*

early diagnosis of Alzheimer's disease., Phys Med Biol, 56 (2011), pp. 6047–6063.

[37] Y.-M. Cheung, *k*-means: A new generalized k-means clustering algorithm*, Pattern Recognition Letters, 24 (2003), pp. 2883–2893.

[38] D. B. Chklovskii, S. Vitaladevuni, and L. K. Scheffer, *Semi-automated reconstruction of neural circuits using electron microscopy.*, Curr Opin Neurobiol, 20 (2010), pp. 667–675.

[39] F. R. Chung, *Spectral graph theory*, vol. 92, American Mathematical Soc., 1997.

[40] Committee on Metagenomics: Challenges and Functional Applications, National Research Council, *The New Science of Metagenomics: Revealing the Secrets of Our Microbial Planet*, The National Academies Press, 2007.

[41] R. C. Craddock, M. P. Milham, and S. M. LaConte, *Predicting intrinsic brain activity.*, Neuroimage, 82 (2013), pp. 127–136.

[42] A. Daducci, S. Gerhard, A. Griffa, A. Lemkaddem, L. Cammoun, X. Gigandet, R. Meuli, P. Hagmann, and J.-P. Thiran, *The connectome mapper: an open-source processing pipeline to map connectomes with MRI.*, PLoS One, 7 (2012), p. e48121.

[43] M. de Jager, K. M. Trujillo, P. Sung, K.-P. Hopfner, J. P. Carney, J. A. Tainer, J. C. Connelly, D. R. F. Leach, R. Kanaar, and C. Wyman, *Differential arrangements of conserved building blocks among homologs of the Rad50/Mre11 DNA repair protein complex.*, J Mol Biol, 339 (2004), pp. 937–949.

[44] S. C. de Lange, M. A. de Reus, and M. P. van den Heuvel, *The-Laplacian spectrum of neural networks.*, Front Comput Neurosci, 7 (2014), p. 189.

[45] S. Delmas, L. Shunburne, H.-P. Ngo, and T. Allers, *Mre11-Rad50 promotes rapid repair of DNA damage in the polyploid archaeon Haloferax volcanii by restraining homologous recombination.*, PLoS Genet, 5 (2009), p. e1000552.

[46] I. S. Dhillon, Y. Guan, and B. Kulis, *Kernel k-means: spectral clustering and normalized cuts*, in Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining, KDD '04, New York, NY, USA, 2004, ACM, pp. 551–556.

[47] J. M. Dick and E. L. Shock, *Calculation of the relative chemical stabilities of proteins as a function of temperature and redox chemistry in a hot spring.*, PLoS One, 6 (2011), p. e22782.

[48] B. Dubois, H. H. Feldman, C. Jacova, S. T. Dekosky, P. Barberger-Gateau, J. Cummings, A. Delacourte, D. Galasko, S. Gauthier, G. Jicha, K. Meguro, J. O'brien, F. Pasquier, P. Robert, M. Rossor, S. Salloway, Y. Stern, P. J. Visser, and P. Scheltens, *Research criteria for the diagnosis of Alzheimer's disease: revising the NINCDS-ADRDA criteria.*, Lancet Neurol, 6 (2007), pp. 734–746.

[49] S. Dymond, N. S. Lawrence, B. T. Dunkley, K. S. L. Yuen, E. C. Hinton, M. R. Dixon, W. M. Cox, A. E. Hoon, A. Munnelly, S. D. Muthukumaraswamy, and K. D. Singh, *Almost winning: induced meg theta power in insula and orbitofrontal cortex increases during gambling near-misses and is associated with bold signal and gambling severity.*, Neuroimage, 91 (2014), pp. 210–219.

[50] S. R. Eddy, *Accelerated profile HMM searches.*, PLoS Comput Biol, 7 (2011), p. e1002195.

[51] R. J. Epstein, *Unblocking blockbusters: using boolean text-mining to optimise clinical trial design and timeline for novel anticancer drugs.*, Cancer Inform, 7 (2009), pp. 231–238.

[52] M. Fenech, *Folate (vitamin B9) and vitamin B12 and their function in the maintenance of nuclear and mitochondrial genome integrity.*, Mutat Res, 733 (2012), pp. 21–33.

[53] M. Ferrer, O. Golyshina, A. Beloqui, and P. N. Golyshin, *Mining enzymes from extreme environments.*, Curr Opin Microbiol, 10 (2007), pp. 207–214.

[54] N. Fierer, J. W. Leff, B. J. Adams, U. N. Nielsen, S. T. Bates, C. L. Lauber, S. Owens, J. A. Gilbert, D. H. Wall, and J. G. Caporaso, *Cross-biome metagenomic analyses of soil microbial communities and their functional attributes.*, Proc Natl Acad Sci U S A, 109 (2012), pp. 21390–21395.

[55] C. Fine, *Neuroscience. his brain, her brain?*, Science, 346 (2014), pp. 915–916.

[56] B. Fischl, *Freesurfer*, Neuroimage, 62 (2012), pp. 774–781.

[57] L. R. Ford and D. R. Fulkerson, *Maximal flow through a network*, Canadian Journal of Mathematics, 8 (1956), pp. 399–404.

[58] C. Galustian and A. G. Dalgleish, *The power of the web in cancer drug discovery and clinical trial design: research without a laboratory?*, Cancer Inform, 9 (2010), pp. 31–35.

[59] M. R. Garey, D. S. Johnson, and L. Stockmeyer, *Some simplified NP-complete graph problems*, Theoretical computer science, 1 (1976), pp. 237–267.

[60] C. Gilbert, *Brain connectivity: revealing the fly visual motion circuit.*, Curr Biol, 23 (2013), pp. R851–R853.

[61] D. Ginsburg, S. Gerhard, J. E. Congote, and R. Pienaar, *Real-time visualization of the connectome in the browser using webgl*, Frontiers in Neuroinformatics, (2011).

[62] E. M. Glass, J. Wilkening, A. Wilke, D. Antonopoulos, and F. Meyer, *Using the metagenomics RAST server (MG-RAST) for an-*

alyzing shotgun metagenomes., Cold Spring Harb Protoc, 2010 (2010), p. pdb.prot5368.

[63] J. P. González-Brenes and J. Mostow, *What and when do students learn? fully data-driven joint estimation of cognitive and student models.*

[64] D. J. Graham, *Routing in the brain.*, Front Comput Neurosci, 8 (2014), p. 44.

[65] H.-Y. Ha, S.-C. Chen, and M. Chen, *Fc-mst: Feature correlation maximum spanning tree for multimedia concept classification*, in Semantic Computing (ICSC), 2015 IEEE International Conference on, IEEE, 2015, pp. 276–283.

[66] P. Hagmann, L. Cammoun, X. Gigandet, R. Meuli, C. J. Honey, V. J. Wedeen, and O. Sporns, *Mapping the structural core of human cerebral cortex.*, PLoS Biol, 6 (2008), p. e159.

[67] P. Hagmann, P. E. Grant, and D. A. Fair, *Mr connectomics: a conceptual framework for studying the developing brain.*, Front Syst Neurosci, 6 (2012), p. 43.

[68] D. J. Hand, H. Mannila, and P. Smyth, *Principles of Data Mining*, MIT Press, 2001.

[69] R. J. Hathaway and J. C. Bezdek, *Nerf c-means: Non-euclidean relational fuzzy clustering*, Pattern Recognition, 27 (1994), pp. 429–437.

[70] J. R. Havig, J. Raymond, D. R. Meyer-Dombard, N. Zolotova, and E. L. Shock, *Merging isotopes and community genomics in a siliceous sinter-depositing hot spring*, Journal of Geophysical Research: Biogeosciences (2005–2012), 116 (2011).

[71] T. Heimo, K. Kaski, and J. Saramäki, *Maximal spanning trees, asset graphs and random matrix denoising in the analysis of dynamics of financial networks*, Physica A: Statistical Mechanics and its Applications, 388 (2009), pp. 145–156.

[72] E. P. Helzner, J. A. Luchsinger, N. Scarmeas, S. Cosentino, A. M. Brickman, M. M. Glymour, and Y. Stern, *Contribution of vascular risk factors to the progression in Alzheimer's disease.*, Arch Neurol, 66 (2009), pp. 343–348.

[73] D. S. Hochbaum, *Approximation algorithms for the set covering and vertex cover problems*, SIAM Journal on Computing, 11 (1982), pp. 555–556.

[74] P. G. Hoel, *Introduction to mathematical statistics.*, John Wiley & Sons, Inc., New York, 5fth ed., 1984.

[75] S. Holm, *A simple sequentially rejective multiple test procedure*, Scandinavian Journal of Statistics, (1979), pp. 65–70.

[76] S. Hoory, N. Linial, and A. Wigderson, *Expander graphs and their applications*, Bulletin of the American Mathematical Society, 43 (2006), pp. 439–561.

[77] K. P. Hopfner, A. Karcher, L. Craig, T. T. Woo, J. P. Carney, and J. A. Tainer, *Structural biochemistry and interaction architecture of the DNA double-strand break repair Mre11 nuclease and Rad50-ATPase.*, Cell, 105 (2001), pp. 473–485.

[78] K. P. Hopfner, A. Karcher, D. S. Shin, L. Craig, L. M. Arthur, J. P. Carney, and J. A. Tainer, *Structural biology of Rad50 ATPase: ATP-driven conformational control in DNA double-strand break repair and the ABC-ATPase superfamily.*, Cell, 101 (2000), pp. 789–800.

[79] L. D. Hurst and A. R. Merchant, *High guanine-cytosine content is not an adaptation to high temperature: a comparative analysis amongst prokaryotes.*, Proc Biol Sci, 268 (2001), pp. 493–497.

[80] D. H. Huson, A. F. Auch, J. Qi, and S. C. Schuster, *MEGAN analysis of metagenomic data.*, Genome Res, 17 (2007), pp. 377–386.

[81] D. H. Huson and S. Mitra, *Introduction to the analysis of environmental sequences: metagenomics with MEGAN.*, Methods Mol Biol, 856 (2012), pp. 415–429.

[82] D. H. Huson, S. Mitra, H.-J. Ruscheweyh, N. Weber, and S. C. Schuster, *Integrative analysis of environmental sequences using MEGAN4.*, Genome Res, 21 (2011), pp. 1552–1560.

[83] M. Ingalhalikar et al., *Sex differences in the structural connectome of the human brain.*, Proc Natl Acad Sci U S A, 111 (2014), p. 823.

[84] M. Ingalhalikar, A. Smith, D. Parker, T. D. Satterthwaite, M. A. Elliott, K. Ruparel, H. Hakonarson, R. E. Gur, R. C. Gur, and R. Verma, *Reply to Joel and Tarrasch: On misreading and shooting the messenger.*, Proc Natl Acad Sci U S A, 111 (2014), p. E638.

[85] G. Ivan, Z. Szabadka, and V. Grolmusz, *Being a binding site: Characterizing residue composition of binding sites on proteins.*, Bioinformation, 2 (2007), pp. 216–221.

[86] C. R. Jack, V. J. Lowe, S. D. Weigand, H. J. Wiste, M. L. Senjem, D. S. Knopman, M. M. Shiung, J. L. Gunter, B. F. Boeve, B. J. Kemp, M. Weiner, R. C. Petersen, and A. D. N. Initiative, *Serial PIB and MRI in normal, mild cognitive impairment and Alzheimer's disease: implications for sequence of pathological events in Alzheimer's disease.*, Brain, 132 (2009), pp. 1355–1365.

[87] A. L. Jacobs and P. Schar, *DNA glycosylases: in DNA repair and beyond.*, Chromosoma, 121 (2012), pp. 1–20.

[88] D. Joel and R. Tarrasch, *On the mis-presentation and misinterpretation of gender-related data: the case of Ingalhalikar's human connectome study.*, Proc Natl Acad Sci U S A, 111 (2014), p. E637.

[89] H. Johansen-Berg, *Human connectomics - what will the future demand?*, Neuroimage, 80 (2013), pp. 541–544.

[90] S. Karlin, J. Mrázek, and A. M. Campbell, *Compositional biases of bacterial genomes and evolutionary implications.*, J Bacteriol, 179 (1997), pp. 3899–3913.

[91] C. Kerepesi, D. Banky, and V. Grolmusz, *AmphoraNet: the webserver implementation of the AMPHORA2 metagenomic workflow suite.*, Gene, 533 (2014), pp. 538–540.

[92] C. Kerepesi, B. Szalkai, and V. Grolmusz, *Visual analysis of the quantitative composition of metagenomic communities: the AmphoraVizu webserver.*, Microb Ecol, (2014).

[93] F. Kiefer, K. Arnold, M. Kunzli, L. Bordoli, and T. Schwede, *The swiss-model repository and associated resources.*, Nucleic Acids Res, 37 (2009), pp. D387–D392.

[94] G. Kirchhoff, *Über die Auflösung der Gleichungen, auf welche man bei der untersuchung der linearen verteilung galvanischer Ströme geführt wird*, Ann. Phys. Chem., 72 (1847).

[95] L. Krause, N. N. Diaz, A. Goesmann, S. Kelley, T. W. Nattkemper, F. Rohwer, R. A. Edwards, and J. Stoye, *Phylogenetic classification of short environmental DNA fragments.*, Nucleic Acids Res, 36 (2008), pp. 2230–2239.

[96] E. L. Lawler, *Combinatorial optimization: networks and matroids*, Courier Dover Publications, 1976.

[97] L. Lovász, *Combinatorial problems and exercises*, American Mathematical Society, 2nd ed., June 2007.

[98] A. Lyra, S. Lahtinen, K. Tiihonen, and A. C. Ouwehand, *Intestinal microbiota and overweight.*, Benef Microbes, 1 (2010), pp. 407–421.

[99] J. B. MacQueen, *Some methods for classification and analysis of multivariate observations*, in Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability, L. M. L. Cam and J. Neyman, eds., University of California Press, 1967, pp. 281–297.

[100] M. Malaguarnera, R. Ferri, R. Bella, G. Alagona, A. Carnemolla, and G. Pennisi, *Homocysteine, vitamin B12 and folate in vascular dementia and in Alzheimer's disease.*, Clin Chem Lab Med, 42 (2004), pp. 1032–1035.

[101] R. N. Mantegna, *Hierarchical structure in financial markets*, The European Physical Journal B-Condensed Matter and Complex Systems, 11 (1999), pp. 193–197.

[102] A. McCaddon, S. Tandy, P. Hudson, R. Gray, G. Davies, D. Hill, and J. Duguid, *Absence of macrocytic anaemia in Alzheimer's disease.*, Clin Lab Haematol, 26 (2004), pp. 259–263.

[103] J. A. McNab, B. L. Edlow, T. Witzel, S. Y. Huang, H. Bhat, K. Heberlein, T. Feiweier, K. Liu, B. Keil, J. Cohen-Adad, M. D. Tisdall, R. D. Folkerth, H. C. Kinney, and L. L. Wald, *The Human Connectome Project and beyond: initial applications of 300 mT/m gradients.*, Neuroimage, 80 (2013), pp. 234–245.

[104] M. M. Mielke, P. P. Zandi, M. Sjogren, D. Gustafson, S. Ostling, B. Steen, and I. Skoog, *High total cholesterol levels in late life associated with a reduced risk of dementia.*, Neurology, 64 (2005), pp. 1689–1695.

[105] E. Mossello, E. Ballini, A. M. Mello, F. Tarantini, D. Simoni, S. Baldasseroni, and N. Marchionni, *Biomarkers of Alzheimer's disease: from central nervous system to periphery?*, Int J Alzheimers Dis, 2011 (2010), p. 342980.

[106] V. Nemeth-Pongracz, O. Barabas, M. Fuxreiter, I. Simon, I. Pichova, M. Rumlova, H. Zabranska, D. Svergun, M. Petoukhov,

V. Harmat, E. Klement, E. Hunyadi-Gulyas, K. F. Medzihradszky, E. Konya, and B. G. Vertessy, *Flexible segments modulate co-folding of dUTPase and nucleocapsid proteins.*, Nucleic Acids Res, 35 (2007), pp. 495–505.

[107] J. Neu, G. Lorca, S. D. K. Kingma, and E. W. Triplett, *The intestinal microbiome: relationship to type 1 diabetes.*, Endocrinol Metab Clin North Am, 39 (2010), pp. 563–571.

[108] S. E. O'Bryant, G. Xiao, R. Barber, J. Reisch, R. Doody, T. Fairchild, P. Adams, S. Waring, R. Diaz-Arrastia, and T. A. R. Consortium, *A serum protein-based algorithm for the detection of Alzheimer disease.*, Arch Neurol, 67 (2010), pp. 1077–1081.

[109] S. E. O'Bryant, G. Xiao, R. Barber, J. Reisch, J. Hall, C. M. Cullum, R. Doody, T. Fairchild, P. Adams, K. Wilhelmsen, R. Diaz-Arrastia, T. A. Research, and C. Consortium, *A blood-based algorithm for the detection of Alzheimer's disease.*, Dement Geriatr Cogn Disord, 32 (2011), pp. 55–62.

[110] F. O'Leary, M. Allman-Farinelli, and S. Samman, *Vitamin B12 status, cognitive decline and dementia: a systematic review of prospective cohort studies.*, Br J Nutr, 108 (2012), pp. 1948–1961.

[111] A. S. Pandit, M. N. Joshi, P. Bhargava, G. N. Ayachit, I. M. Shaikh, Z. M. Saiyed, A. K. Saxena, and S. B. Bagatharia, *Metagenomes from the saline desert of Kutch.*, Genome Announc, 2 (2014).

[112] N. Pierrot, T. D., D. L., D. I., G. P., H. A., T. B., H. LE., M. N., N. F., L. A., D. JB., C. D., B. JP., C. PJ., K.-C. P, and O. JN., *Amyloid precursor protein controls cholesterol turnover needed for neuronal activity*, EMBO Mol Med, 5 (2013), pp. 608–.

[113] M. Prince and J. Jackson, *World Alzheimer Report 2009*, tech. rep., Alzheimer's Disease International, 2009.

[114] H. Prüfer, *Neuer beweis eines satzes über permutationen*, Arch. Math. Phys, 27 (1918), pp. 742–744.

[115] J. Qin, Y. Li, Z. Cai, S. Li, J. Zhu, F. Zhang, S. Liang, W. Zhang, Y. Guan, D. Shen, Y. Peng, D. Zhang, Z. Jie, W. Wu, Y. Qin, W. Xue, J. Li, L. Han, D. Lu, P. Wu, Y. Dai, X. Sun, Z. Li, A. Tang, S. Zhong, X. Li, W. Chen, R. Xu, M. Wang, Q. Feng, M. Gong, J. Yu, Y. Zhang, M. Zhang, T. Hansen, G. Sanchez, J. Raes, G. Falony, S. Okuda, M. Almeida, E. LeChatelier, P. Renault, N. Pons, J.-M. Batto, Z. Zhang, H. Chen, R. Yang, W. Zheng, S. Li, H. Yang, J. Wang, S. D. Ehrlich, R. Nielsen, O. Pedersen, K. Kristiansen, and J. Wang, *A metagenome-wide association study of gut microbiota in type 2 diabetes.*, Nature, 490 (2012), pp. 55–60.

[116] C. Reitz, J. Luchsinger, M.-X. Tang, J. Manly, and R. Mayeux, *Impact of plasma lipids and time on memory performance in healthy elderly without dementia.*, Neurology, 64 (2005), pp. 1378–1383.

[117] C. Reitz, M.-X. Tang, J. Luchsinger, and R. Mayeux, *Relation of plasma lipids to Alzheimer disease and vascular dementia.*, Arch Neurol, 61 (2004), pp. 705–714.

[118] P. Rice, I. Longden, and A. Bleasby, *EMBOSS: the European Molecular Biology Open Software Suite.*, Trends Genet, 16 (2000), pp. 276–277.

[119] J. A. Rogers, D. Polhamus, W. R. Gillespie, K. Ito, K. Romero, R. Qiu, D. Stephenson, M. R. Gastonguay, and B. Corrigan, *Combining patient-level and summary-level data for Alzheimer's disease modeling and simulation: a beta regression meta-analysis.*, J Pharmacokinet Pharmacodyn, 39 (2012), pp. 479–498.

[120] K. Romero, B. Corrigan, C. W. Tornoe, J. V. Gobburu, M. Danhof, W. R. Gillespie, M. R. Gastonguay, B. Meibohm, and

H. Derendorf, *Pharmacometrics as a discipline is entering the "industrialization" phase: standards, automation, knowledge sharing, and training are critical for future success.*, J Clin Pharmacol, 50 (2010), pp. 9S–19S.

[121] K. Romero, M. de Mars, D. Frank, M. Anthony, J. Neville, L. Kirby, K. Smith, and R. L. Woosley, *The coalition against major diseases: developing tools for an integrated drug development process for alzheimer's and parkinson's diseases.*, Clin Pharmacol Ther, 86 (2009), pp. 365–367.

[122] I. Saeed and S. K. Halgamuge, *The oligonucleotide frequency derived error gradient and its application to the binning of metagenome fragments.*, BMC Genomics, 10 Suppl 3 (2009), p. S10.

[123] C. Schmeisser, H. Steele, and W. R. Streit, *Metagenomics, biotechnology with non-culturable microbes.*, Appl Microbiol Biotechnol, 75 (2007), pp. 955–962.

[124] R. Seshadri, S. A. Kravitz, L. Smarr, P. Gilna, and M. Frazier, *CAMERA: a community resource for metagenomics.*, PLoS Biol, 5 (2007), p. e75.

[125] F. Sievers, A. Wilm, D. Dineen, T. J. Gibson, K. Karplus, W. Li, R. Lopez, H. McWilliam, M. Remmert, J. Soding, J. D. Thompson, and D. G. Higgins, *Fast, scalable generation of high-quality protein multiple sequence alignments using clustal omega.*, Mol Syst Biol, 7 (2011), p. 539.

[126] E. L. Sonnhammer, S. R. Eddy, and R. Durbin, *Pfam: a comprehensive database of protein domain families based on seed alignments.*, Proteins, 28 (1997), pp. 405–420.

[127] C. Spearman, *The proof and measurement of association between two things*, The American Journal of Psychology, 15 (1904), pp. 72–101.

[128] H. L. Steele and W. R. Streit, *Metagenomics: advances in ecology and biotechnology.*, FEMS Microbiol Lett, 247 (2005), pp. 105–111.

[129] D.-P. Streitbuerger, H. E. Moller, M. Tittgemeyer, M. Hund-Georgiadis, M. L. Schroeter, and K. Mueller, *Investigating structural brain changes of dehydration using voxel-based morphometry.*, PLoS One, 7 (2012), p. e44195.

[130] N. Sueoka, *On the genetic basis of variation and heterogeneity of DNA base composition.*, Proc Natl Acad Sci U S A, 48 (1962), pp. 582–592.

[131] J. G. Sutcliffe, P. B. Hedlund, E. A. Thomas, F. E. Bloom, and B. S. Hilbush, *Peripheral reduction of beta-amyloid is sufficient to reduce brain beta-amyloid: implications for Alzheimer's disease.*, J Neurosci Res, 89 (2011), pp. 808–814.

[132] W. D. Swingley, D. R. Meyer-Dombard, E. L. Shock, E. B. Alsop, H. D. Falenski, J. R. Havig, and J. Raymond, *Coordinating environmental genomics and geochemistry reveals metabolic transitions in a hot spring ecosystem.*, PLoS One, 7 (2012), p. e38108.

[133] B. Szalkai, *Generalizing k-means for an arbitrary distance matrix*, ArXiv e-prints http://arxiv.org/abs/1303.6001, (2013).

[134] B. Szalkai, *An implementation of the relational k-means algorithm*, CoRR, abs/1304.6899 (2013).

[135] B. Szalkai and V. Grolmusz, *Nucleotide 9-mers characterize the type ii diabetic gut metagenome*, Genomics, (2016), pp. –.

[136] B. Szalkai, V. K. Grolmusz, V. I. Grolmusz, and C. Against Major Diseases, *Identifying Combinatorial Biomarkers by Association Rule Mining in the CAMD Alzheimer's Database*, ArXiv e-prints, (2013).

[137] B. Szalkai, C. Kerepesi, B. Varga, and V. Grolmusz, *The budapest reference connectome server v2.0*, Neuroscience Letters, 595 (2015), pp. 60 – 62.

[138] B. Szalkai, I. Scheer, K. Nagy, B. G. Vértessy, and V. Grolmusz, *The metagenomic telescope*, PLoS ONE, 9 (2014), pp. 1–9.

[139] B. Szalkai, B. Varga, and V. Grolmusz, *The advantage is at the ladies: Brain size bias-compensated graph-theoretical parameters are also better in women's connectomes*, arXiv preprint arXiv:1512.01156, (2015).

[140] B. Szalkai, B. Varga, and V. Grolmusz, *Graph theoretical analysis reveals: Women's brains are better connected than men's*, PLoS ONE, 10 (2015), p. e0130045.

[141] B. Szalkai, B. Varga, and V. Grolmusz, *Graph theoretical analysis reveals: Women's brains are better connected than men's*, arXiv preprint arXiv:1501.00727, (2015).

[142] B. Szalkai, B. Varga, and V. Grolmusz, *Mapping Correlations of Psychological and Connectomical Properties of the Dataset of the Human Connectome Project with the Maximum Spanning Tree Method*, ArXiv e-prints, (2016).

[143] B. Szigeti, P. Gleeson, M. Vella, S. Khayrulin, A. Palyanov, J. Hokanson, M. Currie, M. Cantarelli, G. Idili, and S. Larson, *Openworm: an open-science approach to modelling caenorhabditis elegans*, Frontiers in Computational Neuroscience, 8 (2014).

[144] R. E. Tarjan, *Data structures and network algorithms*, vol. 44 of CBMS-NSF Regional Conference Series in Applied Mathematics, Society for Industrial Applied Mathematics, 1983.

[145] E. K. Towlson, P. E. Vértes, S. E. Ahnert, W. R. Schafer, and E. T. Bullmore, *The rich club of the c. elegans neuronal connectome.*, J Neurosci, 33 (2013), pp. 6380–6387.

[146] Q. Tu, Z. He, and J. Zhou, *Strain/species identification in metagenomes using genome-specific markers.*, Nucleic Acids Res, 42 (2014), p. e67.

[147] K. Turgutalp, O. Ozhan, E. Gok Oguz, A. Yilmaz, M. Horoz, I. Helvaci, and A. Kiykim, *Community-acquired hypernatremia in elderly and very elderly patients admitted to the hospital: clinical characteristics and outcomes.*, Med Sci Monit, 18 (2012), pp. CR729–CR734.

[148] B. Varga, O. Barabas, J. Kovari, J. Toth, E. Hunyadi-Gulyas, E. Klement, K. F. Medzihradszky, F. Tolgyesi, J. Fidy, and B. G. Vertessy, *Active site closure facilitates juxtaposition of reactant atoms for initiation of catalysis by human dUTPase.*, FEBS Lett, 581 (2007), pp. 4783–4788.

[149] B. Varga, O. Barabas, E. Takacs, N. Nagy, P. Nagy, and B. G. Vertessy, *Active site of mycobacterial dUTPase: structural characteristics and a built-in sensor.*, Biochem Biophys Res Commun, 373 (2008), pp. 8–13.

[150] S. Weintraub, S. S. Dikmen, R. K. Heaton, D. S. Tulsky, P. D. Zelazo, P. J. Bauer, N. E. Carlozzi, J. Slotkin, D. Blitz, K. Wallner-Allen, et al., *Cognition assessment using the nih toolbox*, Neurology, 80 (2013), pp. S54–S64.

[151] J. G. White, E. Southgate, J. N. Thomson, and S. Brenner, *The structure of the nervous system of the nematode caenorhabditis elegans*, Philosophical Transactions of the Royal Society of London B: Biological Sciences, 314 (1986), pp. 1–340.

[152] R. A. Whitmer, S. Sidney, J. Selby, S. C. Johnston, and K. Yaffe, *Midlife cardiovascular risk factors and risk of dementia in late life.*, Neurology, 64 (2005), pp. 277–281.

[153] A. Wilke, E. M. Glass, D. Bartels, J. Bischof, D. Braithwaite, M. D'Souza, W. Gerlach, T. Harrison, K. Keegan, H. Matthews, R. Kottmann, T. Paczian, W. Tang, W. L. Trimble, P. Yilmaz, J. Wilkening, N. Desai, and F. Meyer, *A metagenomics portal for a democratized sequencing world.*, Methods Enzymol, 531 (2013), pp. 487–523.

[154] G. J. Williams, S. P. Lees-Miller, and J. A. Tainer, *Mre11-Rad50-Nbs1 conformations and the control of sensing, signaling, and effector responses at DNA double-strand breaks.*, DNA Repair (Amst), 9 (2010), pp. 1299–1306.

[155] R. S. Williams, G. Moncalian, J. S. Williams, Y. Yamada, O. Limbo, D. S. Shin, L. M. Groocock, D. Cahill, C. Hitomi, G. Guenther, D. Moiani, J. P. Carney, P. Russell, and J. A. Tainer, *Mre11 dimers coordinate DNA end bridging and nuclease processing in double-strand-break repair.*, Cell, 135 (2008), pp. 97–109.

[156] S. F. Witelson, H. Beresh, and D. L. Kigar, *Intelligence and brain size in 100 postmortem brains: sex, lateralization and age factors.*, Brain, 129 (2006), pp. 386–398.

[157] T. H. Wonnacott and R. J. Wonnacott, *Introductory statistics*, vol. 19690, Wiley New York, 1972.

[158] M. Wortmann, *Dementia: a global health priority - highlights from an ADI and World Health Organization report.*, Alzheimers Res Ther, 4 (2012), p. 40.

[159] H. Wu, Z. Zhang, S. Hu, and J. Yu, *On the molecular mechanism of GC content variation among eubacterial genomes.*, Biol Direct, 7 (2012), p. 2.

[160] M. Wu and J. A. Eisen, *A simple, fast, and accurate method of phylogenomic inference.*, Genome Biol, 9 (2008), p. R151.

[161] M. Wu and A. J. Scott, *Phylogenomic analysis of bacterial and archaeal sequences with AMPHORA2.*, Bioinformatics, 28 (2012), pp. 1033–1034.

[162] S. Wu and Y. Zhang, *Muster: Improving protein sequence profile-profile alignments by using multiple sources of structure information.*, Proteins, 72 (2008), pp. 547–556.

[163] W. Wu, W. C. Juan, C. R. M. Y. Liang, K. G. Yeoh, J. So, and M. C. M. Chung, *S100A9, GIF and AAT as potential combinatorial biomarkers in gastric cancer diagnosis and prognosis.*, Proteomics Clin Appl, 6 (2012), pp. 152–162.

[164] X. Wu, C. Ma, L. Han, M. Nawaz, F. Gao, X. Zhang, P. Yu, C. Zhao, L. Li, A. Zhou, J. Wang, J. E. Moore, B. C. Millar, and J. Xu, *Molecular characterisation of the faecal microbiota in patients with type II diabetes.*, Curr Microbiol, 61 (2010), pp. 69–78.

[165] W. Xie, F. Wang, L. Guo, Z. Chen, S. M. Sievert, J. Meng, G. Huang, Y. Li, Q. Yan, S. Wu, X. Wang, S. Chen, G. He, X. Xiao, and A. Xu, *Comparative metagenomics of microbial communities inhabiting deep-sea hydrothermal vent chimneys with contrasting chemistries.*, ISME J, 5 (2011), pp. 414–426.

[166] C. Yook, S. Druckmann, and J. Kim, *Mapping mammalian synaptic connectivity.*, Cell Mol Life Sci, 70 (2013), pp. 4747–4757.

[167] D. Zambon, M. Quintana, P. Mata, R. Alonso, J. Benavent, F. Cruz-Sanchez, J. Gich, M. Pocov, F. Civeira, S. Capurro, D. Bachman, K. Sambamurti, J. Nicholas, and M. A. Pappolla, *Higher incidence of mild cognitive impairment in familial hypercholesterolemia.*, Am J Med, 123 (2010), pp. 267–274.

[168] J. Živković, M. Mitrović, and B. Tadić, *Correlation patterns in gene expressions along the cell cycle of yeast*, no. 207 in Studies in Computational Intelligence, Springer, 2009.

[1]ADATLAP
a doktori értekezés nyilvánosságra hozatalához

I. A doktori értekezés adatai
A szerző neve: **Szalkai Balázs**
MTMT-azonosító: **10054370**
A doktori értekezés címe és alcíme: **Algoritmikus kérdések a bioinformatika területén**
DOI-azonosító[2]: **10.15476/ELTE.2018.077**
A doktori iskola neve: **ELTE Informatika Doktori Iskola**
A doktori iskolán belüli doktori program neve: **Információs rendszerek**
A témavezető neve és tudományos fokozata: **Dr. Grolmusz Vince egyetemi tanár**
A témavezető munkahelye: **ELTE Számítógéptudományi Tanszék**

II. Nyilatkozatok
1. A doktori értekezés szerzőjeként[3]
a) hozzájárulok, hogy a doktori fokozat megszerzését követően a doktori értekezésem és a tézisek nyilvánosságra kerüljenek az ELTE Digitális Intézményi Tudástárban. Felhatalmazom az Informatika Doktori Iskola hivatalának ügyintézőjét, Kulcsár Adinát, hogy az értekezést és a téziseket feltöltse az ELTE Digitális Intézményi Tudástárba, és ennek során kitöltse a feltöltéshez szükséges nyilatkozatokat.
b) kérem, hogy a mellékelt kérelemben részletezett szabadalmi, illetőleg oltalmi bejelentés közzétételéig a doktori értekezést ne bocsássák nyilvánosságra az Egyetemi Könyvtárban és az ELTE Digitális Intézményi Tudástárban;
c) kérem, hogy a nemzetbiztonsági okból minősített adatot tartalmazó doktori értekezést a minősítés (*dátum*)-ig tartó időtartama alatt ne bocsássák nyilvánosságra az Egyetemi Könyvtárban és az ELTE Digitális Intézményi Tudástárban;
d) kérem, hogy a mű kiadására vonatkozó mellékelt kiadó szerződésre tekintettel a doktori értekezést a könyv megjelenéséig ne bocsássák nyilvánosságra az Egyetemi Könyvtárban, és az ELTE Digitális Intézményi Tudástárban csak a könyv bibliográfiai adatait tegyék közzé. Ha a könyv a fokozatszerzést követőn egy évig nem jelenik meg, hozzájárulok, hogy a doktori értekezésem és a tézisek nyilvánosságra kerüljenek az Egyetemi Könyvtárban és az ELTE Digitális Intézményi Tudástárban.
2. A doktori értekezés szerzőjeként kijelentem, hogy
a) az ELTE Digitális Intézményi Tudástárba feltöltendő doktori értekezés és a tézisek saját eredeti, önálló szellemi munkám és legjobb tudomásom szerint nem sértem vele senki szerzői jogait;
b) a doktori értekezés és a tézisek nyomtatott változatai és az elektronikus adathordozón benyújtott tartalmak (szöveg és ábrák) mindenben megegyeznek.
3. A doktori értekezés szerzőjeként hozzájárulok a doktori értekezés és a tézisek szövegének plágiumkereső adatbázisba helyezéséhez és plágiumellenőrző vizsgálatok lefuttatásához.

Kelt: **Budapest, 2018. április 19.**

a doktori értekezés szerzőjének aláírása

[1] Beiktatta az Egyetemi Doktori Szabályzat módosításáról szóló CXXXIX/2014. (VI. 30.) Szen. sz. határozat. Hatályos: 2014. VII.1. napjától.
[2] A kari hivatal ügyintézője tölti ki.
[3] A megfelelő szöveg aláhúzandó.